PH.D. THESIS

Heavy and light traffic regimes for M|G|infinity traffic models

by Konstantinos P. Tsoukatos
Advisor: Armand M. Makowski

CSHCN Ph.D. 99-3
(ISR Ph.D. 99-6)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1999		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Heavy and light traffic regimes for M|G|infinity traffic models	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Army Research Laboratory,2800 Powder Mill Road,Adelphi,MD,20783**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**155**	

ABSTRACT

Title of Dissertation: HEAVY AND LIGHT TRAFFIC REGIMES FOR $M|G|\infty$ TRAFFIC MODELS

Konstantinos P. Tsoukatos, Doctor of Philosophy, 1999

Dissertation directed by: Professor Armand M. Makowski
Department of Electrical Engineering

The $M|G|\infty$ busy server process provides a class of structural models for communication network traffic. In this dissertation, we study the asymptotic behavior of a network multiplexer, modeled as a discrete–time queue, driven by an $M|G|\infty$ correlated arrival stream. The asymptotic regimes considered here are those of heavy and light traffic. In heavy traffic, we show that the arising limits are described in terms of the classical Brownian motion and the α–stable Lévy motion, under short– and long–range dependence, respectively. Salient features are then effectively captured by the exponential distribution and the Mittag–Leffler special function. In light traffic, the analysis reveals the effect of two aspects of the $M|G|\infty$ process, i.e., the session duration distribution G and the gradual nature of the arrivals, as opposed to the instantaneous inputs of a standard $GI|GI|1$ queue. We exploit these asymptotic results to construct interpolation approximations for system quantities of interest, applicable to all traffic intensities.

HEAVY AND LIGHT TRAFFIC REGIMES FOR $M|G|\infty$ TRAFFIC MODELS

by

Konstantinos P. Tsoukatos

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1999

Advisory Committee:

Professor Armand M. Makowski, Chairperson/Advisor
Professor Prakash Narayan
Associate Professor Adrian Papamarcou
Associate Professor Leandros Tassiulas
Professor A. Udaya Shankar, University Representative

DEDICATION

To my parents and to my teachers

ACKNOWLEDGMENTS

I am indebted to my advisor Professor Armand M. Makowski who initiated me in heavy tailed stochastic models and provided me with financial support throughout the past three years. In the course of this work I greatly benefited from his ideas, suggestions, criticism, and attention to detail. His meticulous review of the various iterations of the manuscript helped me eliminate several mistakes and substantially improve my writing style.

I would like to thank Professors Prakash Narayan, Adrian Papamarcou, Leandros Tassiulas, and A. Udaya Shankar for kindly consenting to serve in my PhD dissertation committee.

I also thank Dr. Mark S. Squillante, who hosted me as a summer intern at IBM T.J. Watson Research Center, and gave me ample opportunity to study topics some of which found their way into this thesis.

This research was conducted in the Electrical Engineering Department and Institute for Systems Research at the University of Maryland, College Park, and was funded through NSF Grant NSFD CDR-88-03012, NASA Grant NAGW277S and the Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0002. I also acknowledge the support of Eugenides Foundation, Athens, Greece, in the form of a Scholarship that enabled me to pursue graduate studies in the United States.

During my stay at Maryland I met several people, mostly Greeks but others as well, who made these years much more enjoyable, on and off campus. For their companionship and for the pleasant memories that I will relish in the future I acknowledge them with great pleasure.

Lastly, and most importantly, I would like to express my gratitude to my parents $\Pi\alpha\nu\alpha\gamma\iota\acute{\omega}\tau\eta s$ and $\Sigma\tau\upsilon\lambda\iota\alpha\nu\acute{\eta}$, for instilling in me the perseverance without which this work would have never been completed.

TABLE OF CONTENTS

List of Tables **viii**

List of Figures **ix**

1 Introduction **1**

1.1 From on/off to $M|G|\infty$ sources 3

1.2 Summary and discussion 4

1.3 Self–similarity, stable distributions and regular variation 7

2 $M|G|\infty$ and related models **14**

2.1 The $M|G|\infty$ arrival processes 14

2.1.1 Definitions and basic properties 14

2.1.2 Second order self–similarity 18

2.2 The queueing system 20

2.3 Instantaneous inputs 21

2.3.1 A Markov chain of the $M|G|1$ type 22

2.3.2 Case $c = 1$: A solution by generating function 24

2.3.3 Case $c = 1$: An equivalent representation 25

2.3.4 Case $c = 1$: Idle and busy periods 26

2.3.5 Case $c = 1/m$ $(m = 1, 2, \ldots)$ 28

2.3.6 Poisson inputs: Stochastic comparisons 29

3 Heavy traffic: Lévy motion limits **34**

3.1 Introduction 34

3.2 The heavy traffic regime 35

3.3 The main heavy traffic results 38

3.4 Consequences and comments 41

3.4.1 Queue size 42

3.4.2 On selecting the heavy traffic scaling 44

3.5 Outline of proof and preliminary results 47

3.6 Proofs of Theorems 3.3.1–3.3.3 50

3.7 Proofs of Propositions 3.5.1, 3.5.2 and 3.5.3 56

3.8 A proof of Proposition 3.6.1 62

4 Light traffic limits **70**

4.1 Introduction 70

4.2 Reiman–Simon theory 71

4.2.1 Preliminaries 71

4.2.2 Light traffic derivatives 72

4.2.3 Case $c \geq 1$ 75

4.2.4 Case $c = 1$ 77

4.2.5 A heavy–light traffic relationship 82

4.3 Gradual inputs 84

4.3.1 Stationary version 85

4.3.2 Case $c \leq 1$: A stochastic comparison 86

4.3.3 Case $c \leq 1$: Expected queue size $\mathbf{E}[q_\infty]$ 88

4.3.4 Case $c = 1$: Determination of $\mathbf{P}[q_\infty = 0]$ 89
4.3.5 Case $c = 1$: Short– vs long–range dependence 93
4.4 Proof of Lemma 4.2.1 96

5 Interpolation approximations **100**
5.1 Introduction 100
5.2 Summary of asymptotics 101
5.3 Heavy–light traffic interpolations 104
5.3.1 Tail probability approximations 104
5.3.2 Moment approximations 106
5.3.3 Long–range dependence 107
5.4 Numerical results 109
5.5 On the Mittag–Leffler distribution 121

6 Conclusions **125**

Appendix **128**

A Asymptotic invertibility of regularly varying functions **128**

B Stochastic orderings **133**

Bibliography **136**

LIST OF TABLES

5.1 $\mathbf{P}_\lambda[q_\infty > 0]$ for deterministic session duration $\sigma = 3$. 110

5.2 $\mathbf{P}_\lambda[q_\infty > 4]$ for deterministic session duration $\sigma = 3$. 111

5.3 Utilization $\rho = 0.2$; $\sigma \sim$ uniform$(1, 5)$. 112

5.4 Utilization $\rho = 0.8$; $\sigma \sim$ uniform$(1, 5)$. 112

LIST OF FIGURES

4.1 Queue length evolution under the event $\{t_1, t_2; k_1, k_2\}$ 79

4.2 Values of $q_0(\{t_1, t_2; k_1, k_2\})$ when $k_1 \geq k_2$. 80

4.3 Values of $q_0(\{t_1, t_2; k_1, k_2\})$ when $k_1 < k_2$. 80

4.4 Calculation of $\widetilde{\psi}(k_1, k_2)$; example for $k_1 = 2$, $k_2 = 5$. 96

4.5 Values of $\mathbf{1}\left[q_0 > b\right](\{t_1, t_2; k_1, k_2\})$ for $k_1 < k_2$. 99

5.1 Geometric $\gamma = 0.8$ session duration. 113

5.2 Geometric $\gamma = 2/3$ session duration; release rate $c = 4$. 114

5.3 Pareto $\alpha = 1.5$ session duration. 115

5.4 Pareto $\alpha = 1.7$ session duration. 115

5.5 Truncated Pareto $\alpha = 1.7$, $N = 50$. 117

5.6 SRD vs LRD approximation for truncated Pareto session duration. 119

5.7 Truncated Pareto $\alpha = 1.7$, $N = 1000$. 120

5.8 Integration path 122

5.9 Mittag–Leffler distributions $E_\nu(-x^\nu)$ and spectral densities $f_\nu(x)$. . 124

Chapter 1

Introduction

With high resolution traffic measurements from a variety of networking applications becoming widely available, there has been renewed interest in understanding the statistical nature of emerging network traffic. Large data sets, obtained from Ethernet LANs, VBR video sequences, ftp, telnet and WWW applications in WANs, have been extensively studied and results point to the conclusion that real traffic is very bursty, exhibiting great variability over extended periods of time that are much longer than previously expected. As these dependencies may have a pronounced effect on performance [40], they should be taken into account when modeling network traffic for buffer and link provisioning, or for evaluating scheduling policies.

In the recent literature on traffic modeling such persistent correlations are often reported to be best captured by stochastic processes that are long–range dependent [4, 12, 22, 23, 36, 48]. Roughly speaking, this happens when the data stream displays correlations which span multiple time scales, and which, despite being individually small, decay in such a slow hyperbolic–like manner as to be considered non–summable. In other cases (e.g., the studies of VBR video traces [30, 34]) sample autocorrelation functions are found to conform with a more general subex-

ponential decay. In any case, the observed dependencies cannot be exclusively attributed to classical Markovian models, with bounded exponential moments. A consensus seems to be emerging to the fact that non–traditional stochastic models should be considered; these will most likely play an increasing role in capturing the dynamics of traffic that networks are expected to carry in the near future.

Going beyond the statistical findings mentioned above, an interesting line of current research focuses on quantifying the consequences of high variability and dependence on network performance. The initial experimental work in [18] indicates that the impact of long–range dependencies is adverse and significant, yet evidence to the contrary also exists [24, 38, 52], suggesting that in many practical situations queueing measures are not seriously affected. In addition, the traffic measurement studies have generated interest in queueing systems with correlated arrival processes. Few analytical results are currently available for queues with long–range or subexponentially dependent arrivals. These include the fractional Brownian motion model of Norros [43], fractional Gaussian noise [1], the popular independent on/off source model with Pareto activity periods [9, 28] and, more recently, the multiplexed on/off sources [29]. In all these cases buffer overflow probabilities display a slow non–exponential decay; this is in sharp contrast with the exponential tails that typically characterize queues with short–range dependent Markovian inputs. Moreover, a closer examination confirms that classification in terms of the short vs. long–range dependent nature of the traffic is often insufficient: Both within the short and the long–range dependent regime further details matter, and vastly different queueing behaviors arise.

1.1 From on/off to $M|G|\infty$ sources

In this dissertation, we consider the class of discrete–time $M|G|\infty$ input processes. An $M|G|\infty$ input process is understood as the busy server process of a discrete–time infinite server system fed by a discrete–time Poisson process of rate λ (customers/slot) and with generic service time σ. Such $M|G|\infty$ processes can account both for short and long–range dependent behaviors, with the correlation patterns controlled through σ [Proposition 2.1.1]. Furthermore, asymptotic self–similarity arises when σ is Pareto–like, i.e., has a regularly varying tail of the form (3.9). $M|G|\infty$ processes have already been used by Paxson and Floyd to successfully model WAN traffic [48]. However, perhaps the most convincing justification supporting their use as plausible traffic models is provided by the following limiting result [37]: Consider M identical and independent sources, with alternating independent emission and silence periods. Assume that during its "on" periods each source generates information at a constant rate of one unit per time slot, while during the "off" periods it remains inactive. We alternatively view these "on" periods as corresponding to information sessions. Denote by σ the duration (in slots) of the generic information session and allow the duration of the generic "off" period, denoted by T_{off}, to depend on the number of sources, i.e., $T_{\text{off}} = T_{\text{off}}(M)$. The resulting total session arrival rate $\lambda(M)$ is given by

$$\lambda(M) = \frac{M}{\mathbf{E}[\sigma] + \mathbf{E}[T_{\text{off}}(M)]}.$$

Let the number of sources M go to infinity, while simultaneously reducing the number of sessions of an individual source, so that the aggregate session arrival rate remains finite. This can be achieved by selecting $T_{\text{off}}(M)$ such that $\mathbf{E}[T_{\text{off}}(M)] = M/\lambda$ for some $\lambda > 0$, in which case $\lim_{M\to\infty} \lambda(M) = \lambda$. Consider now the process

which, in each time slot, records the total number of newly initiated "on" periods. This converges, as M goes to infinity, to a discrete–time Poisson process with session arrival rate λ (sessions/slot). Since σ is the generic session duration random variable (rv), we readily identify the $M|G|\infty$ busy server process with parameters (λ, σ) as the limiting process counting the number of active sessions at any given time slot. Hence, the class of $M|G|\infty$ processes is one that naturally arises from a Poisson superposition scheme of infinitely many simpler on/off sources.

1.2 Summary and discussion

As shown in [14, 39, 46, 47], $M|G|\infty$ processes induce a wide variety of asymptotic behaviors for the buffer probabilities at a multiplexer with constant release rate. In particular, when σ has a regularly varying tail – the $M|G|\infty$ process is now asymptotically self–similar – the buffer asymptotics are hyperbolic in nature, in stark contrast with the Weibullian tails induced by fractional Gaussian noise (or fractional Brownian motion) [43]. A key contribution of this dissertation is to elucidate the noted difference in buffer asymptotics between $M|G|\infty$ and fractional Gaussian noise inputs by further exploring this discrepancy in the heavy traffic regime. One might expect that, with asymptotically identical correlation patterns, both models necessarily have a heavy traffic characterization in terms of fractional Brownian motion, in very much the same manner that different short–range dependent models eventually collapse to a single description involving Brownian motion. However, this turns out not to be the case.

In Chapter 3 we show that, under short–range dependence, the class of $M|G|\infty$ inputs belongs to the domain of attraction of the standard Brownian motion, as expected. However, under long–range dependence, with σ belonging to the domain

of attraction of a non–normal stable law, the $M|G|\infty$ process is not attracted to a fractional Brownian motion, but instead to a non–Gaussian, α–stable Lévy motion which is $1/\alpha$ self–similar. As a consequence, the distribution of the heavy traffic queue length is given by a Mittag–Leffler function, thus displaying not a Weibullian, but a Pareto tail, with power $1-\alpha$ [Theorem 3.4.3]. These results undescore the fundamentally different nature of the long–range dependent $M|G|\infty$ process (when compared to fractional Gaussian noise), and also point to the fact that fractional Brownian motion does not necessarily play for long–range dependence the same key role that standard Brownian motion assumes under short–range dependence. Within long–range dependence, there seems to be a choice for distinct modeling possibilities, and it is not at all difficult to find rather simple, potentially useful traffic models that are attracted to non–Gaussian limits.

In Chapter 4 we shift attention to the light traffic regime. This refers to the limiting situation where the traffic intensity approaches zero. Noting that the $M|G|\infty$ process is Poisson driven, we apply the Reiman–Simon theory [49, 50, 51] to obtain information in the form of derivatives of system quantities with respect to the intensity of the driving Poisson process, when this intensity tends to zero [Propositions 4.2.3, 4.2.4]. In addition, when the "on"–state rate of each constituent on/off source exceeds the multiplexer release rate explicit expressions for the expected queue size become available. These results quantify the differences between the gradual $M|G|\infty$ inputs and the point arrivals of a classical $GI|GI|1$ queue, and suggest a classification of the light traffic behavior of the buffer content distribution in terms of the short– vs long– range dependent property of the $M|G|\infty$ process [Corollaries 4.3.1, 4.3.2]. However, in light traffic further subcases arise depending on comparisons of the on/off source "on"–state rate and the server

capacity. Thus, in general, the correlation function and the short– vs long–range dependence property of the $M|G|\infty$ inputs are not the sole factors that impact performance.

Work on queueing analysis under long–range dependence appears to have been initiated by Norros [43], where the presence of fractional Brownian motion is postulated. This line of inquiry is further pursued in [58], while in [9] Brichet et al. show how fractional Brownian motion can arise from a Gaussian superposition scheme of infinitely many on/off sources with heavy tailed on/off periods. In the limiting setup of [9] the sources are "small", i.e., the peak rate of the individual source becomes infinitely small in comparison with the multiplexer capacity. In view of the fact that $M|G|\infty$ processes arise from a different superposition scheme of infinitely many on/off sources [37], where the peak on/off source rate is "large", i.e., remains comparable to the link capacity, it is not too surprising that these lead to a different heavy traffic limit involving Lévy motions. More recently, in [8], a heavy traffic limit of this type, giving rise to a Mittag–Leffler function, is obtained in the standard $GI|GI|1$ queueing setup, for the case where the service time distribution is heavy tailed. Heavy traffic results similar and related to the ones given here have also been reported in [33], where only convergence of finite dimensional distributions is announced. The conclusions discussed here were obtained independently, and were summarized in the conference paper [61].

The asymptotic characterization of the queue size distribution in the heavy and light traffic regimes is exploited in Chapter 5. By suitably interpolating between the two extremes we derive approximations to the queue size distribution, applicable to all traffic intensities. For some common choices for the session duration distribution G the approximants assume a simple final form. The accuracy

of the proposed expressions as well as pitfalls of this technique are discussed via several numerical examples [Section 5.4]. In Chapters 2 and 4 we make occasional detours and study simpler queueing systems; these help us obtain exact results [Propositions 4.3.2, 4.3.3] and establish stochastic comparisons that can provide bounds whenever exact expressions are not available. In Appendices A and B we have summarized several needed facts concerning functions of regular variation and stochastic orderings.

A few words about the notation adopted here. We use $\Longrightarrow_r$ to denote weak convergence [5], and $\xrightarrow{P}_r$ to denote convergence in probability (with r going to infinity). We write $f(x) \sim g(x)$ $(x \to \infty)$ when $\lim_{x\to\infty} f(x)/g(x) = 1$. Equality in distribution is denoted by $=_{st}$, inequality in the strong, convex and increasing convex stochastic ordering sense are denoted by $\leq_{st}$, $\leq_{cx}$ and $\leq_{icx}$, respectively.

1.3 Self–similarity, stable distributions and regular variation

This section provides a quick tour into some background concepts which recur throughout the dissertation. The material presented here, and much more, can be found in [3, 7, 16, 20, 53]. We start with a definition of long–range dependence:

Definition 1.3.1 *The* $\mathbb{R}$*–valued wide sense stationary process* $\{Y_k,\ k = 0, \pm 1, \ldots\}$ *is said to be long–range dependent if*

$$\sum_{k=1}^{\infty} |\mathrm{cov}\,[Y_k, Y_0]\,| = \infty \tag{1.1}$$

and short–range dependent otherwise.

We proceed with self–similarity. Roughly speaking, a structure is "self–similar" if it appears the same on any scale, large or small. The term self–similar and the definition below are due to Mandelbrot [41].

Definition 1.3.2 *The* $\mathbb{R}$*–valued process* $\{X(t),\ t \in \mathbb{R}\}$ *is (strictly) self–similar with index (or Hurst parameter)* $H > 0$ *if for all* $a > 0$ *the finite-dimensional distributions of* $\{X(at),\ t \in \mathbb{R}\}$ *are identical to the finite-dimensional distributions of* $\{a^H X(t),\ t \in \mathbb{R}\}$*, i.e., if for any* $n = 1, 2, \ldots,\ t_1, t_2, \ldots, t_n$ *in* $\mathbb{R}$*, and* $a > 0$*,*

$$(X(at_1), X(at_2), \ldots, X(at_n)) =_{st} (a^H X(t_1), a^H X(t_2), \ldots, a^H X(t_n)).$$

From this definition we see that, in the context of stochastic processes, self–similarity is tantamount to scale invariance of the finite–dimensional distributions, but not necessarily of the sample paths.

Definition 1.3.3 *We say that the* $\mathbb{R}$*–valued process* $\{X(t),\ t \in \mathbb{R}\}$ *is* H*-sssi if it is strictly self–similar with index* $H > 0$ *and has stationary increments.*

Among the H-sssi processes the Gaussian one is the most prominent; this is in part due to the fact that it has been widely applied in the context of long–range dependence.

Definition 1.3.4 *A* H*-sssi Gaussian process with index* $0 < H \leq 1$ *is called fractional Brownian motion and is denoted by* $\{B_H(t),\ t \in \mathbb{R}\}$*. It is called standard fractional Brownian motion if* $\mathrm{var}[B_H(1)] = 1$*.*

Proposition 1.3.1 *With* $0 < H \leq 1$*, the fractional Brownian motion* $\{B_H(t),\ t \in \mathbb{R}\}$ *has the following properties:*

(a) $B_H(0) = 0$ *a.s.*

(b) *Its covariance function is given by*

$$\operatorname{cov}[B_H(t_1), B_H(t_2)] = \frac{1}{2}\left\{|t_1|^{2H} + |t_2|^{2H} - |t_1 - t_2|^{2H}\right\} \operatorname{var}[B_H(1)], \quad t_1, t_2 \in \mathbb{R}\,.$$

(c) *When $0 < H < 1$, we have $\mathbf{E}\,[B_H(t)] = 0$ for all t in $\mathbb{R}$.*

(d) *When $H = 1$, we have $B_1(t) = tB_1(1)$ a.s. for all t in $\mathbb{R}$.*

In the case $H = 1/2$ Proposition 1.3.1(b) reads

$$\operatorname{cov}[B_{1/2}(t_1), B_{1/2}(t_2)] = \begin{cases} \operatorname{var}\left[B_{1/2}(1)\right] \min(t_1, t_2) & \text{if } t_1 t_2 > 0 \\ 0 & \text{if } t_1 t_2 \leq 0 \end{cases}$$

and $\{B_{1/2}(t),\ t \in \mathbb{R}\}$ is classical Brownian motion.

Consider now the increment process $\{Z_H(n),\ n = 0, \pm 1, \ldots\}$ associated with $\{B_H(t),\ t \in \mathbb{R}\}$ and defined by

$$Z_H(n) := B_H(n+1) - B_H(n), \quad n = 0, \pm 1, \ldots. \tag{1.2}$$

This Gaussian sequence is stationary, because fractional Brownian motion $\{B_H(t),\ t \in \mathbb{R}\}$ has stationary increments.

Definition 1.3.5 *The stationary process $\{Z_H(n),\ n = 0, \pm 1, \ldots\}$ of (1.2) is called fractional Gaussian noise. It is called standard fractional Gaussian noise if* $\operatorname{var}[Z_H(1)] = 1$.

From Definition 1.3.5 and Proposition 1.3.1(b) it follows that the covariance function $r_H(n) := \operatorname{cov}\left[Z_H(n+1), Z_H(1)\right]$ of the fractional Gaussian noise process $\{Z_H(n),\ n = 0, \pm 1, \ldots\}$ is given by

$$r_H(n) = \frac{1}{2}\left\{|n+1|^{2H} - 2|n|^{2H} + |n-1|^{2H}\right\} \operatorname{var}\left[Z_H(1)\right], \quad n = 0, \pm 1, \ldots. \tag{1.3}$$

If $H = 1/2$ then $r_H(n) = 0$ for $n \neq 0$, in which case $\{Z_{1/2}(n),\ n = 0, \pm 1, \ldots\}$ forms a sequence of i.i.d. Gaussian rvs and fractional Gaussian noise reduces to the familiar white Gaussian noise. For $H \neq 1/2$ it follows from (1.3) that

$$r_H(n) \sim \mathrm{var}\,[Z_H(1)]\, H(2H-1)\, n^{2H-2} \quad (n \to \infty), \tag{1.4}$$

so that $r_H(n)$ behaves like a power function. Note that $\lim_{n\to\infty} r_H(n) = 0$ for all $0 < H < 1$, however, for $1/2 < H < 1$ the covariance function $r_H(n)$ decays so slowly as $n \to \infty$ that the corresponding sum (1.1) diverges. Thus, in the case $1/2 < H < 1$ the fractional Gaussian noise process $\{Z_H(n),\ n = 0, \pm 1, \ldots\}$ is long–range dependent.

We next present another class of processes which also contains the classical Brownian motion as a special case. First, recall the family of stable distributions.

Definition 1.3.6 *A non-degenerate rv X is said to have a stable distribution if for any positive numbers c_1 and c_2 there is a positive number $a(c_1, c_2)$ and a real number $b(c_1, c_2)$ such that*

$$c_1 X_1 + c_2 X_2 =_{st} a(c_1, c_2) X + b(c_1, c_2),$$

where X_1 and X_2 are i.i.d copies of X.

The characteristic functions of all stable distributions were discovered by Lévy (1924):

Proposition 1.3.2 *A rv X has a stable distribution if and only if its characteristic function is of the form*

$$\mathbf{E}\,[\exp(i\theta X)] = \exp\left\{i\mu\theta - \delta^\alpha |\theta|^\alpha \left(1 - i\beta\,\mathrm{sgn}(\theta)\ z(\theta, \alpha)\right)\right\}, \quad \theta \in \mathbb{R} \tag{1.5}$$

where μ *is a real constant,* $\delta > 0$*,* α *in* $(0, 2]$*,* β *in* $[-1, 1]$ *and*

$$z(\theta, \alpha) := \begin{cases} \tan\left(\frac{\pi\alpha}{2}\right) & \text{if } \alpha \neq 1, \\ -\frac{2}{\pi}\ln|\theta| & \text{if } \alpha = 1. \end{cases}$$

Stable distributions are the only possible limit distributions for normalized and centered sums of i.i.d. rvs, and in that sense they generalize the Gaussian distribution, which is obtained by setting $\alpha = 2$ in (1.5). When $0 < \alpha < 2$ methods for generating deviates from stable laws are available but, with a few exceptions, closed forms expressions for stable densities are not known. However, series expansions and the tail behavior of stable distributions are known.

Let $S_\alpha(\delta, \beta, \mu)$ denote the generic stable rv distributed according to (1.5). The next two results can be found in the monograph by Samorodnitsky and Taqqu [53, pp. 16, 18].

Proposition 1.3.3 *If* $X =_{st} S_\alpha(\delta, \beta, \mu)$ *with* $0 < \alpha < 2$*, then*

$$\lim_{x \to +\infty} x^\alpha \mathbf{P}[X > x] = K_\alpha \frac{1+\beta}{2}\, \delta^\alpha \tag{1.6}$$

and

$$\lim_{x \to +\infty} x^\alpha \mathbf{P}[X < -x] = K_\alpha \frac{1-\beta}{2}\, \delta^\alpha \tag{1.7}$$

where $K_\alpha := \left(\int_0^\infty x^{-\alpha} \sin x\, dx\right)^{-1}$.

As a consequence, we have

Proposition 1.3.4 *If* $X =_{st} S_\alpha(\delta, \beta, \mu)$ *with* $0 < \alpha < 2$*, then*

$$\mathbf{E}[|X|^p] < \infty \quad \text{for} \quad 0 < p < \alpha,$$

and

$$\mathbf{E}[|X|^p] = \infty \quad \text{for} \quad p \geq \alpha.$$

In particular, for $1 < \alpha < 2$ *it holds that*

$$\text{var}\,[X] = \infty \quad \textit{and} \quad \mathbf{E}\,[|X|] < \infty.$$

Because of this infinite p^{th} moment property for $p > \alpha$, stable distributions are candidates for modeling phenomena with high variability.

Let us now consider the following stochastic process:

Definition 1.3.7 *With* α *in* $(0, 2]$, *the* $\mathbb{R}$*–valued process* $\{L_\alpha(t),\ t \geq 0\}$ *is called (standard)* α*–stable Lévy motion if*

(a) $L_\alpha(0) = 0 \quad a.s.,$

(b) $\{L_\alpha(t),\ t \geq 0\}$ *has independent increments, and*

(c) $L_\alpha(t) - L_\alpha(s) =_{st} S_\alpha((t-s)^{1/\alpha}, \beta, 0)$ *for all* $0 \leq s < t < \infty$, *for some* β *in* $[-1, 1]$.

Clearly, the 2–stable motion $\{L_2(t),\ t \geq 0\}$ is simply the Brownian motion. Moreover, using (1.5) it can be verified that for all $c > 0$ the processes $\{L_\alpha(ct),\ t \geq 0\}$ and $\{c^{1/\alpha} L_\alpha(t),\ t \geq 0\}$ have the same finite–dimensional distributions, and an α–stable Lévy motion $\{L_\alpha(t),\ t \geq 0\}$ is self–similar with index $H = 1/\alpha$ (unless $\alpha = 1$ and $\beta \neq 0$).

In view of the power–like tail behaviors encountered in (1.4), (1.6) and (1.7), it is appropriate to review the notion of regular variation. The definition below can be interpreted as introducing a class of generalized power functions:

Definition 1.3.8 *A Lebesgue measurable function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *is said to be regularly varying (at infinity) with index* ρ *in* $\mathbb{R}$ *if*

$$\lim_{x \to +\infty} \frac{f(xy)}{f(x)} = y^\rho, \quad y > 0. \tag{1.8}$$

If $\rho = 0$ *in (1.8) then* f *is called slowly varying.*

From (1.8) it follows that if $f : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying with index ρ then it can be written as

$$f(x) = x^\rho \; h(x), \quad x > 0,$$

where the function $h : \mathbb{R}_+ \to \mathbb{R}_+$ is slowly varying.

Work on functions of regular variation was initiated by Karamata (1930). Later, regular variation and in particular its relevance in probability was popularized by Feller [20]; an authoritative treatment of the subject is found in the monograph [7]. The following theorem due to Lamperti [35] (see also [7, p. 356], [3, p. 50]) provides a connection between regular variation and self–similarity: Self–similar processes are exactly those that arise from probability limit theorems where a process is centered and rescaled.

We use $\stackrel{f.d.d.}{\Longrightarrow}$ to denote convergence of finite–dimensional distributions.

Theorem 1.3.1 *Suppose that the* $\mathbb{R}$*–valued process* $\{X(t),\ t \in \mathbb{R}\}$ *is such that, with suitably chosen mappings* $f, g : \mathbb{R} \to \mathbb{R}$

$$\left\{ \frac{X(ut) - g(u)}{f(u)},\ t \in \mathbb{R} \right\} \stackrel{f.d.d.}{\Longrightarrow} \{Y(t),\ t \in \mathbb{R}\} \quad (u \to \infty) \tag{1.9}$$

for some $\mathbb{R}$*–valued process* $\{Y(t),\ t \in \mathbb{R}\}$ *with non-degenerate* $Y(1)$*. Then* $\{Y(t),\ t \in \mathbb{R}\}$ *is strictly self–similar, and all self–similar processes arise in this way. Moreover,* f *is regularly varying with index* H*, where* H *is the Hurst parameter of the limiting self–similar process* $\{Y(t),\ t \in \mathbb{R}\}$*.*

The fact that the norming functions f appearing in (1.9) are necessarily regularly varying shows that regular variation is intrinsically present in limit theorems of probability.

Chapter 2

$M|G|\infty$ and related models

In this chapter we introduce a class of traffic models based on the $M|G|\infty$ busy server process. In addition, we discuss simpler related queueing models, with which stochastic comparisons will be sought.

2.1 The $M|G|\infty$ arrival processes

We start by presenting the $M|G|\infty$ arrival processes, together with the assumptions and notation that will be used throughout. Several key properties concerning this class of processes are stated here without proof; additional details are given in [11, 47].

2.1.1 Definitions and basic properties

Consider a population of infinitely many information sources, operating in discrete–time. Sources can be in one of two states, active or idle. During time slot $[n, n+1)$, $n = 0, 1, \ldots$, β_{n+1} new sources become active. Source j, $j = 1, \ldots, \beta_{n+1}$, begins generating information by the start of slot $[n+1, n+2)$, its activity period has duration $\sigma_{n+1,j}$ (in number of slots). While active, each source generates

information at a constant rate of one information unit (packet) per time slot. After its activity period expires, each source switches off permanently, never to generate packets again. Let b_n denote the number of active sources, or equivalently, the number of packets generated by the active sources at the beginning of time slot $[n, n+1)$. If initially (i.e., at time $n = 0$) there were already b active sources, we denote by $\sigma_{0,j}$ the residual activity duration (in time slots) for the j^{th} active source, $j = 1, \ldots, b$.

For each $n = 0, 1, \ldots$ we have the decomposition

$$b_n = b_n^{(0)} + b_n^{(a)} \tag{2.1}$$

where the rvs $b_n^{(0)}$ and $b_n^{(a)}$ describe the contributions to the number of active sources at the beginning of slot $[n, n+1)$ from the sources already active at $n = 0$ and from subsequent activations in slots $[k, k+1)$, $k = 1, 2, \ldots, n$, respectively. We readily check that

$$b_n^{(0)} = \sum_{j=1}^{b} \mathbf{1}\left[\sigma_{0,j} > n\right] \quad \text{and} \quad b_n^{(a)} = \sum_{k=1}^{n} \sum_{j=1}^{\beta_k} \mathbf{1}\left[\sigma_{k,j} > n-k\right]. \tag{2.2}$$

The rv $b_n^{(a)}$ can also be interpreted as the number of active sources at the beginning of slot $[n, n+1)$ given that all sources were silent at time $n = 0$. On the other hand, to obtain a stationary process $\{b_n,\ n = 0, 1, \ldots\}$, the rv $b_n^{(0)}$ should be specified as the number of active sources at time $n = 0$ given that the sources have been operating since time $n = -\infty$. This requirement dictates the appropriate distributional assumptions on the rvs b and $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$. With these considerations in mind, we now record a set of assumptions enforced throughout.

Assumption (A) *The $\mathbb{N}$–valued rvs b, $\{\beta_{n+1},\ n = 0, 1, \ldots\}$, $\{\sigma_{n,j},\ n = 1, 2, \ldots;\ j = 1, 2, \ldots\}$ and $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ are defined on a common probability space*

$(\Omega, \mathcal{F}, \mathbf{P})$ *and satisfy the following:*

(i) These rvs are mutually independent.

(ii) The rv b is a Poisson rv with parameter $\lambda \mathbf{E}[\sigma]$.

(iii) The rvs $\{\beta_{n+1},\ n = 0, 1, \ldots\}$ *are i.i.d. Poisson rvs with parameter* $\lambda > 0$.

(iv) The rvs $\{\sigma_{n,j},\ n = 1, \ldots;\ j = 1, 2, \ldots\}$ *are i.i.d. with distribution function G on* $\{1, 2, \ldots\}$. *Let* σ *denote the generic* $\mathbb{N}$*–valued rv distributed according to G. We assume that* $\mathbf{E}[\sigma] < \infty$.

(v) The rvs $\{\sigma_{0,j},\ j = 1, 2, \ldots\}$ *are i.i.d.* $\mathbb{N}$*–valued rvs distributed according to the forward recurrence time distribution* $\widehat{G}$ *associated with G, i.e., if* $\widehat{\sigma}$ *denotes a generic* $\mathbb{N}$*–valued rv distributed according to* $\widehat{G}$, *then*

$$\widehat{g}_n := \mathbf{P}[\widehat{\sigma} = n] = \frac{\mathbf{P}[\sigma \geq n]}{\mathbf{E}[\sigma]}, \quad n = 1, 2, \ldots. \tag{2.3}$$

The proposition below summarizes the properties of the resulting process $\{b_n,\ n = 0, 1, \ldots\}$ and is a consequence of Assumption (A), (2.1) and (2.2) [47].

Proposition 2.1.1 *The process* $\{b_n,\ n = 0, 1, \ldots\}$ *is a (strictly) stationary ergodic process with the following properties:*

(a) *For each* $n = 0, 1, \ldots$, *the rv* b_n *is a Poisson rv with parameter* $\lambda \mathbf{E}[\sigma]$;

(b) *Its covariance function is given by*

$$\operatorname{cov}[b_{n+j}, b_n] = \lambda \mathbf{E}\left[(\sigma - j)^+\right] = \lambda \mathbf{E}[\sigma]\, \mathbf{P}[\widehat{\sigma} > j], \quad n, j = 0, 1, \ldots;$$

(c) *Its index of dispersion of counts (IDC) is given by*

$$\mathrm{IDC} := \sum_{j=0}^{\infty} \operatorname{cov}[b_{n+j}, b_n] = \lambda \mathbf{E}[\sigma] \sum_{j=0}^{\infty} \mathbf{P}[\widehat{\sigma} > j] = \frac{\lambda}{2} \mathbf{E}[\sigma(\sigma + 1)] \tag{2.4}$$

and the process is short–range dependent (i.e., IDC finite) if and only if $\mathbf{E}\left[\sigma^2\right]$ *is finite.*

From part (b) above it is clear that the sequence $\{b_n,\ n = 0, 1, \ldots\}$ exhibits some form of positive dependence. In fact, as mentioned in [45], the rvs $\{b_n,\ n = 0, 1, \ldots\}$ are strongly positively correlated, in a sense that can be made precise by using the notion of association [19]:

Proposition 2.1.2 *The rvs* $\{b_n,\ n = 0, 1, \ldots\}$ *are associated, in that for any* $n = 0, 1, \ldots$ *and any pair of non–decreasing mappings* $f, g : \mathbb{N}^{n+1} \to \mathbb{R}$ *we have*

$$\mathbf{E}\left[f(b_0, \ldots, b_n)g(b_0, \ldots, b_n)\right] \geq \mathbf{E}\left[f(b_0, \ldots, b_n)\right]\mathbf{E}\left[g(b_0, \ldots, b_n)\right] \tag{2.5}$$

provided the expectations exist and are finite.

In summary, the process $\{b_n,\ n = 0, 1, \ldots\}$ results from discrete–time Poisson arrivals of information sessions, where the generic session duration rv σ is distributed according to the pmf G and the packet generation rate of an ongoing session is one packet per time slot. It is fully characterized by a pair (λ, G), with λ the Poisson arrival rate (per slot). Under Assumption (A) the sequence $\{b_n,\ n = 0, 1, \ldots\}$ can be identified as the stationary busy server process of a discrete–time $M|G|\infty$ queue; for this reason the packet arrival process $\{b_n,\ n = 0, 1, \ldots\}$ is henceforth referred to as the $M|G|\infty$ input process. From Propositions 2.1.2 and 2.1.1(b) we see that $\{b_n,\ n = 0, 1, \ldots\}$ is an associated process, whose positive correlation structure is completely determined by the distribution of $\widehat{\sigma}$ (and thus of σ). In many cases the inverse is also true, i.e., it is possible to extract $M|G|\infty$ model parameters to match a given autocorrelation function.

Proposition 2.1.3 *An* $\mathbb{R}_+$*–valued sequence* $\{\phi(n),\ n = 0, 1, \ldots\}$ *is the autocorrelation function of an* $M|G|\infty$ *process* (λ, σ) *if and only if the mapping* $n \to \phi(n)$

is decreasing and integer–convex with $\phi(0) = 1 > \phi(1)$ *and* $\lim_{n\to\infty} \phi(n) = 0$*, in which case the corresponding distribution of* σ *is given by*

$$\mathbf{P}\left[\sigma > n\right] = \frac{\phi(n) - \phi(n+1)}{1 - \phi(1)}, \qquad n = 0, 1, \ldots$$

Based on this property, $M|G|\infty$ processes have been used to model VBR video traffic in [34].

2.1.2 Second order self–similarity

In [11] Cox observed that when G is a Pareto distribution with parameter α, $1 < \alpha < 2$, the $M|G|\infty$ busy server process has the so-called second order asymptotic self–similarity property. That is, its correlation structure is asymptotically that of the increments of a strictly self–similar process. The covariance function of standard fractional Gaussian noise is given by (1.3). Using (1.3) and Proposition 2.1.3 we find that if the activity rv σ is distributed according to

$$\mathbf{P}\left[\sigma > n\right] = \frac{|n-1|^{2H} - 3|n|^{2H} + 3|n+1|^{2H} - |n+2|^{2H}}{4(1 - 2^{2H-2})}, \quad n = 0, 1, \ldots, \quad (2.6)$$

with $1/2 < H < 1$, then the corresponding $M|G|\infty$ input process has the same correlation function (1.3) as a fractional Gaussian noise process with Hurst parameter H. This already provides a point of contact between the $M|G|\infty$ process and the increments of a strictly self–similar process. Clearly, the particular distribution (2.6) achieving this match is too restrictive. It is relaxed as follows: For each $m = 1, 2, \ldots$, we introduce the process $\{b_n^{(m)}, n = 0, 1, \ldots\}$ defined by

$$b_n^{(m)} := \frac{1}{m} \sum_{k=0}^{m-1} b_{mn+k}, \quad n = 1, 2, \ldots, \quad (2.7)$$

so that $\{b_n^{(m)}, n = 0, 1, \ldots\}$ is also a stationary process, obtained from $\{b_n,\ n = 0, 1, \ldots\}$ by averaging over blocks of size m. Denote its covariance function by

$$r^{(m)}(k) := \mathrm{cov}[b_n^{(m)}, b_{n+k}^{(m)}], \quad k = 0, 1, \ldots.$$

We say that the original process $\{b_n,\ n = 0, 1, \ldots\}$ is asymptotically second–order self–similar if the correlation function of $\{b_n^{(m)},\ n = 0, 1, \ldots\}$ tends, as the block size m goes to infinity, to the correlation function of fractional Gaussian noise, i.e., if for each lag $k = 1, 2, \ldots$, we have

$$\lim_{m\to\infty} \frac{r^{(m)}(k)}{r^{(m)}(0)} = \frac{r_H(k)}{r_H(0)} \tag{2.8}$$

where $r_H(k)$ is given by (1.3). Noting that

$$r^{(m)}(0) = \frac{1}{m}\left(\mathrm{var}[b_0] + 2\sum_{n=1}^{m}\left(1 - \frac{n}{m}\right)\mathrm{cov}[b_n, b_0]\right) \tag{2.9}$$

it was shown in [37] that the $M|G|\infty$ process is asymptotically second–order self–similar with parameter H, $1/2 < H < 1$, if the tail of σ is regularly varying with index $-(3 - 2H)$, i.e.,

$$\mathbf{P}\left[\sigma > n\right] = n^{-(3-2H)}h(n), \quad n = 1, 2, \ldots, \tag{2.10}$$

for some slowly varying function $h : \mathbb{R}_+ \to \mathbb{R}_+$. The specific distribution (2.6) is simply one instance of (2.10). Whenever $1/2 < H < 1$, the tail behavior (2.10) implies $\mathbf{E}\left[\sigma^2\right] = \infty$, and, by Proposition 2.1.1(b), the corresponding second–order asymptotically self–similar $M|G|\infty$ process is also long–range dependent.

The study of queueing systems in heavy traffic typically involves operations such as accumulating over time and rescaling. For this reason we expect that a heavy traffic analysis of a queue with $M|G|\infty$ arrivals will provide a natural way to further explore the connection with self–similar processes discussed in this section, and address this problem in detail in Chapter 3.

2.2 The queueing system

We now feed the $M|G|\infty$ arrival stream $\{b_n,\ n = 0, 1, \ldots\}$ presented in Section 2.1.1 into a discrete–time single server queue with infinite buffer capacity. Such a queueing system routinely serves as a model for a network multiplexer: If q_n denotes the number of packets remaining in the multiplexer buffer by the end of slot $[n-1, n)$, and the multiplexer output link can transmit c packets/slot, then the buffer content sequence $\{q_n,\ n = 0, 1, \ldots\}$ evolves according to the Lindley recursion

$$q_0 = q; \quad q_{n+1} = [q_n + b_{n+1} - c]^+, \quad n = 0, 1, \ldots \tag{2.11}$$

for some initial buffer content $q \geq 0$. To identify conditions ensuring existence of a finite stationary version of $\{q_n,\ n = 0, 1, \ldots\}$ and to determine its properties, we appeal to established results on recursions of the form (2.11). We introduce the partial sums $\{s_n,\ n = 0, 1, \ldots\}$ defined by

$$s_0 := 0; \quad s_n := \sum_{j=1}^{n} b_j, \quad n = 1, 2, \ldots \tag{2.12}$$

and specialize the results from [2, 54] to the present setup.

Proposition 2.2.1 *If $\lambda \mathbf{E}[\sigma] < c$, then the Lindley recursion (2.11) is termed stable and the following statements hold:*

(a) *There are infinitely many n such that $q_n = 0$.*

(b) *The sequence $\{q_n,\ n = 0, 1, \ldots\}$ of (2.11) and the sequence $\{q_n^0,\ n = 0, 1, \ldots\}$ constructed from (2.11) with $q_0 = 0$ strongly couple, in the sense that*

$$q_n^0 = q_n, \qquad n \geq m(q), \tag{2.13}$$

where $m(q) := min\{n = 0, 1, \ldots : q_n = 0\} < \infty$ (by part (a)).

(c) *The convergence $q_n \Longrightarrow_n q_\infty$ takes place, where the stationary $\mathbb{R}_+$–valued*

rv q_∞ is a.s. finite and given by

$$q_\infty =_{st} \sup\{s_n - nc;\ n = 0, 1, \ldots\}. \tag{2.14}$$

Part (a) above is (1.2.5) in [2, p. 71] (or Lemma 6.1(3) in [54]) and (b) is implied by Lemma 6.1(4) in [54]. Weak convergence in (c) is a consequence of (4.2.6) and Remark 4.1.1 in [2]. Expression (2.14) for the stationary rv q_∞ follows from (2.2.3) of [2].

From parts (b) and (c) of Proposition 2.2.1 we see that if $\lambda \mathbf{E}[\sigma] < c$, then weak convergence to the stationary version q_∞ takes place for any initial condition q; furthermore the distribution of q_∞ does not depend on q. It thus suffices to restrict attention to the choice $q = 0$ and we implicitly do so from now on whenever we refer to (2.11). In that case the system is initially empty and the output to the Lindley recursion admits an equivalent representation given by

$$q_0 = 0; \quad q_n = s_n - nc - \inf\left(s_j - jc,\ j = 0, 1, \ldots, n\right), \quad n = 1, 2, \ldots \tag{2.15}$$

where the partial sums $\{s_n,\ n = 1, 2, \ldots\}$ are defined by (2.12). This is useful for establishing heavy and light traffic limit theorems.

2.3 Instantaneous inputs

By "instantaneous" inputs we refer to the situation where each arriving session brings all of its workload to the system in one time slot, immediately upon arrival. These inputs are to be contrasted with the gradual $M|G|\infty$ inputs, where arriving work is spread over the entire duration of a session. Such instantaneous arrivals are represented by the $\mathbb{N}$–valued sequence of i.i.d. rvs $\{u_n,\ n = 0, 1, \ldots\}$ given by

$$u_{n+1} := \sum_{i=1}^{\beta_{n+1}} \sigma_{n+1,i}, \quad n = 0, 1, \ldots, \tag{2.16}$$

where the families of i.i.d. rvs $\{\beta_{n+1},\ n=0,1,\ldots\}$ and $\{\sigma_{n+1,i},\ n=0,1,\ldots,\ i=1,2,\ldots\}$ are as in Section 2.1.1. These arrivals are also characterized by the pair (λ, G) and we use u to denote the generic rv for the i.i.d. sequence $\{u_n,\ n=0,1,\ldots\}$.

We offer the instantaneous inputs $\{u_n,\ n=0,1,\ldots\}$ to the same multiplexer with constant release rate c. Assuming that the queue is initially empty, we write the corresponding Lindley recursion for the queue length sequence $\{q_n^{(u)},\ n=0,1,\ldots\}$ as

$$q_0^{(u)}=0; \qquad q_{n+1}^{(u)}=[q_n^{(u)}+u_{n+1}-c]^+, \quad n=0,1,\ldots. \tag{2.17}$$

If $\mathbf{E}[u]=\lambda\mathbf{E}[\sigma]<c$ the system is stable and the convergence $q_n^{(u)} \Longrightarrow_n q_\infty^{(u)}$ takes place for some $\mathbb{R}_+$–valued rv $q_\infty^{(u)}$.

Owing to the independence of the rvs $\{u_n,\ n=0,1,\ldots\}$ recursion (2.17) can, at least in principle, be handled by standard generating function techniques. The details of this approach are given in Section 2.3.1.

2.3.1 A Markov chain of the $M|G|1$ type

Consider a Markov chain on $\{0,1,\ldots\}$, whose transition probability matrix P is of the $M|G|1$ type and is given by

$$P=\left\|\begin{array}{cccc} f_0+f_1 & f_2 & f_3 & \ldots \\ f_0 & f_1 & f_2 & \ldots \\ 0 & f_0 & f_1 & \ldots \\ \vdots & & & \end{array}\right\| \tag{2.18}$$

Assume that the probability vector $(f_0, f_1, \ldots)$ satisfies $\sum_{i=0}^{\infty} i f_i < 1$, in which case the Markov chain is positive recurrent. Denote by $(\pi_0, \pi_1, \ldots)$ the steady state

probability vector associated with P, and set

$$\Pi(z) := \sum_{i=0}^{\infty} \pi_i z^i \quad \text{and} \quad F(z) := \sum_{i=0}^{\infty} f_i z^i, \qquad z \in D, \tag{2.19}$$

where $D = \{s \in \mathbb{C} : |s| < 1\}$ is the unit disk in the complex plane $\mathbb{C}$. The vector $(\pi_0, \pi_1, \ldots)$ satisfies

$$\pi_0 = (f_0 + f_1)\pi_0 + f_0\pi_1 \tag{2.20}$$

and

$$\pi_k = \sum_{l=0}^{k+1} f_{k-l+1}\pi_l, \qquad k = 1, 2, \ldots. \tag{2.21}$$

We multiply (2.20) by z and, for each $k = 1, 2, \ldots$, the k^{th} equation in (2.21) by z^k. Adding up, invoking definitions (2.19) and manipulating we find

$$z\Pi(z) = (z-1)f_0\pi_0 + (f_0 + f_1 z)\Pi(z) + (F(z) - f_0 - f_1 z)\Pi(z),$$

so that

$$\Pi(z) = \frac{(z-1)f_0\pi_0}{z - F(z)}, \quad z \in D. \tag{2.22}$$

We compute the limit as $z \to 1$ by applying l' Hospital's rule on the right-hand side in relation (2.22). Since $\Pi(1) = 1$, we get

$$\pi_0 = \frac{1 - F'(1)}{f_0}. \tag{2.23}$$

and inserting (2.23) back in (2.22) yields

$$\Pi(z) = \frac{(1-z)(1-F'(1))}{F(z) - z}, \quad z \in D. \tag{2.24}$$

To calculate $\Pi'(1) = \sum_{i=0}^{\infty} i\pi_i$ we differentiate (2.24) and apply l' Hospital's rule twice, so as to conclude that

$$\Pi'(1) = \frac{F''(1)}{2(1 - F'(1))}. \tag{2.25}$$

2.3.2 Case $c = 1$: A solution by generating function

Set

$$Q^{(u)}(z) := \mathbf{E}\left[z^{q_\infty^{(u)}}\right] \quad \text{and} \quad U(z) := \mathbf{E}\left[z^u\right], \quad z \in D. \tag{2.26}$$

When the multiplexer release rate is $c = 1$ the sequence $\{q_n^{(u)},\ n = 0, 1, \ldots\}$ is a Markov chain on $\{0, 1, \ldots\}$. With the notation of Section 2.3.1, its transition matrix is of the form (2.18) with

$$f_i := \mathbf{P}\left[u = i\right], \qquad i = 0, 1, \ldots. \tag{2.27}$$

Under (2.27), we have the identification

$$F(z) = U(z), \qquad \Pi(z) = Q^{(u)}(z), \qquad z \in D \tag{2.28}$$

and

$$F'(1) = \mathbf{E}\left[u\right], \qquad \Pi'(1) = \mathbf{E}\left[q_\infty^{(u)}\right]. \tag{2.29}$$

As we plan to specialize the results of Section 2.3.1 to the inputs $\{u_n,\ n = 0, 1, \ldots\}$, given by (2.16), we note that

$$U(z) = \exp\left(\lambda(\mathbf{E}\left[z^\sigma\right] - 1)\right) \quad \text{and} \quad U''(1) = \lambda\mathbf{E}\left[\sigma(\sigma - 1)\right] + \lambda^2\mathbf{E}\left[\sigma\right]^2. \tag{2.30}$$

Thus, because of (2.28) and (2.29), relations (2.23), (2.24) and (2.25) imply

$$\mathbf{P}\left[q_\infty^{(u)} = 0\right] = (1 - \lambda\mathbf{E}\left[\sigma\right])e^\lambda \tag{2.31}$$

$$Q^{(u)}(z) = \frac{(1 - z)(1 - \lambda\mathbf{E}\left[\sigma\right])}{\exp\left(\lambda(\mathbf{E}\left[z^\sigma\right] - 1)\right) - z}, \quad z \in D \tag{2.32}$$

and

$$\mathbf{E}\left[q_\infty^{(u)}\right] = \frac{\lambda(\lambda\mathbf{E}\left[\sigma\right]^2 + \mathbf{E}\left[\sigma(\sigma - 1)\right])}{2(1 - \lambda\mathbf{E}\left[\sigma\right])}. \tag{2.33}$$

2.3.3 Case $c = 1$: An equivalent representation

We now derive a representation of the stationary queue size $q_\infty^{(u)}$ in terms of the forward recurrence times of the input sequence $\{u_n,\ n = 1, 2, \ldots\}$. To do this we introduce the sequence of i.i.d. rvs $\{\widehat{u}_n,\ n = 1, 2, \ldots\}$ with generic rv $\widehat{u}$ whose distribution is given by

$$\mathbf{P}\left[\widehat{u} = 0\right] = 0; \qquad \mathbf{P}\left[\widehat{u} = n\right] = \frac{1}{\mathbf{E}\left[u\right]}\,\mathbf{P}\left[u \geq n\right], \quad n = 1, 2, \ldots. \tag{2.34}$$

The corresponding generating function is given by

$$\begin{aligned}
\widehat{U}(z) &:= \sum_{n=1}^{\infty} \mathbf{P}\left[\widehat{u} = n\right] z^n \\
&= \frac{1}{\mathbf{E}\left[u\right]} \sum_{n=1}^{\infty} \mathbf{P}\left[u \geq n\right] z^n \\
&= \frac{1}{\mathbf{E}\left[u\right]} \sum_{n=1}^{\infty} z^n \sum_{k=n}^{\infty} \mathbf{P}\left[u = k\right] \\
&= \frac{1}{\mathbf{E}\left[u\right]} \sum_{k=1}^{\infty} \frac{z^{k+1} - z}{z - 1}\,\mathbf{P}\left[u = k\right] \\
&= \frac{z}{\mathbf{E}\left[u\right]}\,\frac{1 - U(z)}{1 - z}, \quad z \in D.
\end{aligned} \tag{2.35}$$

Under the stability condition $\mathbf{E}\left[u\right] < 1$, relations (2.24) (with the identification (2.28) and (2.29)) read

$$\begin{aligned}
Q^{(u)}(z) &= \frac{(1 - z)(1 - \mathbf{E}\left[u\right])}{U(z) - z} \\
&= (1 - \mathbf{E}\left[u\right]) \left(1 - \frac{1 - U(z)}{1 - z}\right)^{-1}, \quad z \in D.
\end{aligned} \tag{2.36}$$

Therefore, with the help of (2.35), we obtain

$$Q^{(u)}(z) = \frac{1 - \mathbf{E}\left[u\right]}{1 - \mathbf{E}\left[u\right]\mathbf{E}\left[z^{\widehat{u}-1}\right]}, \quad z \in D, \tag{2.37}$$

where we note that the rv $\widehat{u}-1$ is non-negative because of (2.34). Upon rewriting (2.37) as

$$Q^{(u)}(z) = (1 - \mathbf{E}[u]) \sum_{n=0}^{\infty} \mathbf{E}[u]^n \, \mathbf{E}\left[z^{\widehat{u}-1}\right]^n, \quad z \in D,$$

we get the representation

$$q_{\infty}^{(u)} =_{st} \sum_{n=1}^{\nu} (\widehat{u}_n - 1), \tag{2.38}$$

where the rv ν is independent of $\{\widehat{u}_n,\ n = 1, 2, \ldots\}$ and geometrically distributed with parameter $\mathbf{E}[u]$ according to

$$\mathbf{P}[\nu = n] = \mathbf{E}[u]^n (1 - \mathbf{E}[u]), \quad n = 0, 1, \ldots, \tag{2.39}$$

(with the convention that empty sums in (2.38) have value zero). The stationary queue size $q_{\infty}^{(u)}$ is thus expressed as a geometric sum of the forward recurrence times $\{\widehat{u}_n,\ n = 1, 2, \ldots\}$ associated with the input sequence $\{u_n,\ n = 1, 2, \ldots\}$.

2.3.4 Case $c = 1$: Idle and busy periods

We now take one more look at recursion (2.17); this will prove useful in the developments of Chapter 4. We view the queue size sequence $\{q_n^{(u)},\ n = 0, 1, \ldots\}$ as evolving in a series of independent regenerative cycles, alternating between zero and positive values. If the queue is initially empty then, for each $n = 1, 2, \ldots$, the n^{th} cycle consists of an idle period followed by a busy period, with respective lenghts $I_n^{(u)}$ and $B_n^{(u)}$, (expressed in time slots). We say that a time slot is part of a busy period if the queue length at the beginning of the time slot is positive. If the queue length at the left slot boundary is zero, the slot is considered to belong to an idle period. That is, for each $n = 1, 2, \ldots$, the family of i.i.d. pairs of rvs $\{(I_n^{(u)}, B_n^{(u)}),\ n = 1, 2, \ldots\}$ associated with (2.17) are recursively defined by

$$I_n^{(u)} := \inf\{t = 0, 1, \ldots : q^{(u)}_{\sum_{i=1}^{n-1}(I_i+B_i)+t} > 0\} \tag{2.40}$$

and

$$B_n^{(u)} := \inf\{t = 0, 1, \ldots : q^{(u)}_{\sum_{i=1}^{n-1}(I_i+B_i)+I_n+t} = 0\}, \tag{2.41}$$

with the convention that empty sums are zero.

A clarification is needed as the terms "idle" and "busy" are slightly abused here. For example, it is possible that, when the queue length is zero at consecutive instants, say t and t+1, this occurs because of a single arriving packet in $[t, t+1)$ which was served by the end of the time slot. Such a slot is considered to belong to an "idle" period, despite the fact that the server was busy serving the arriving packet. Thus "idle" and "busy" are defined here in reference to queue content, and not to server activity. For lack of better terminology, we shall continue to use "idle" and "busy" in forthcoming arguments, referring to definitions (2.40) and (2.41) to resolve any confusion.

Next, denote by $(I^{(u)}, B^{(u)})$ the generic idle and busy period pair associated with $\{(I_n^{(u)}, B_n^{(u)}),\ n = 1, 2, \ldots\}$. By the Renewal–Reward Theorem we can alternatively express $\mathbf{P}\left[q_\infty^{(u)} = 0\right]$ (which has already been evaluated in (2.31)) as

$$\mathbf{P}\left[q_\infty^{(u)} = 0\right] = \frac{\mathbf{E}\left[I^{(u)}\right]}{\mathbf{E}\left[I^{(u)}\right] + \mathbf{E}\left[B^{(u)}\right]}. \tag{2.42}$$

To obtain the distribution of the idle period $I^{(u)}$ let

$$\eta := \mathbf{P}\left[\beta = 0\right] + \mathbf{P}\left[\beta = 1\right]\mathbf{P}\left[\sigma = 1\right] \tag{2.43}$$

denote the probability that at most one unit of work arrives during a time slot. When $c = 1$, the definitions (2.40) and (2.41) imply that $I^{(u)}$ is geometric with parameter η, i.e.,

$$\begin{aligned} \mathbf{P}\left[I^{(u)} = k\right] &= \mathbf{P}\left[q_1 = \ldots = q_{k-1} = 0, q_k > 0\right] \\ &= \eta^{k-1}(1-\eta), \qquad k = 1, 2, \ldots, \end{aligned} \tag{2.44}$$

so that

$$\mathbf{E}\left[I^{(u)}\right] = \frac{1}{1-\eta}. \tag{2.45}$$

2.3.5 Case $c = 1/m \quad (m = 1, 2, \ldots)$

The arguments presented in Section 2.3.2 can be easily extended to address the following situation: Fix some integer $m = 1, 2, \ldots$ and consider the queue length sequence $\{q_n^{(u)},\ n = 0, 1, \ldots\}$ resulting from the recursion (2.17) with multiplexer release rate $c = 1/m$. In this case $\{q_n^{(u)},\ n = 0, 1, \ldots\}$ is a Markov chain on the lattice $\{0, 1/m, 2/m, \ldots\}$. The transition probability matrix is again of the form (2.18) but this time, instead of (2.27), it holds that

$$f_i := \begin{cases} \mathbf{P}\left[u = \dfrac{i}{m}\right] & \text{if } \ i = 0 \pmod m \\ 0 & \text{if } \ i \neq 0 \pmod m. \end{cases} \tag{2.46}$$

In place of (2.28) we now have the identification

$$F(z) = U(z^m) \quad \text{and} \quad \Pi(z) = Q^{(u)}(z^m), \quad z \in D. \tag{2.47}$$

Consequently,

$$F'(1) = m\mathbf{E}\left[u\right], \quad F''(1) = m^2 U''(1) + m(m-1)U'(1) \tag{2.48}$$

and

$$\Pi'(1) = m\mathbf{E}\left[q_\infty^{(u)}\right]. \tag{2.49}$$

As the analysis leading to (2.23) and (2.25) still applies, we use (2.48) in (2.23) to obtain

$$\mathbf{P}\left[q_\infty^{(u)} = 0\right] = (1 - \lambda m\mathbf{E}\left[\sigma\right])e^\lambda, \tag{2.50}$$

and (2.30), (2.48) and (2.49) in (2.25) to collect

$$\mathbf{E}\left[q_\infty^{(u)}\right] = \frac{\lambda(\lambda m\mathbf{E}\left[\sigma\right]^2 + m\mathbf{E}\left[\sigma^2\right] - \mathbf{E}\left[\sigma\right])}{2(1 - \lambda m\mathbf{E}\left[\sigma\right])}. \tag{2.51}$$

2.3.6 Poisson inputs: Stochastic comparisons

The temporal correlations in the $M|G|\infty$ arrival process are expected to have an adverse effect on queueing performance. Insight to this effect can be obtained from a comparison with the situation where these correlations are altogether eliminated, while maintaining the same Poisson marginal distribution. To that end we consider the companion sequence $\{q_n^{(\xi)},\ n = 0, 1, \ldots\}$ evolving according to

$$q_0^{(\xi)} = 0; \quad q_{n+1}^{(\xi)} = [q_n + \xi_{n+1} - c]^+, \quad n = 0, 1, \ldots \tag{2.52}$$

where the rvs $\{\xi_n,\ n = 1, 2, \ldots\}$ form a sequence of i.i.d. Poisson rvs with parameter $\lambda \mathbf{E}[\sigma]$. Clearly, these independent Poisson arrivals fall in the category of the instantaneous inputs of (2.16), where the pair (λ, σ) is replaced by $(\lambda \mathbf{E}[\sigma], 1)$. Under the stability condition $\lambda \mathbf{E}[\sigma] < c$, convergence to the stationary $\mathbb{R}_+$–valued rv $q_\infty^{(\xi)}$ takes place. Then, for $c = 1/m$, with $m = 1, 2, \ldots$, we can use (2.50) and (2.51) to obtain

$$\mathbf{P}\left[q_\infty^{(\xi)} = 0\right] = (1 - \lambda m \mathbf{E}[\sigma]) e^{\lambda \mathbf{E}[\sigma]} \tag{2.53}$$

and

$$\mathbf{E}\left[q_\infty^{(\xi)}\right] = \frac{\lambda \mathbf{E}[\sigma]\,(\lambda m \mathbf{E}[\sigma] + m - 1)}{2(1 - \lambda m \mathbf{E}[\sigma])}. \tag{2.54}$$

Noting the inequalities

$$\mathbf{P}\left[q_\infty^{(\xi)} = 0\right] \geq \mathbf{P}\left[q_\infty^{(u)} = 0\right] \quad \text{and} \quad \mathbf{E}\left[q_\infty^{(\xi)}\right] \leq \mathbf{E}\left[q_\infty^{(u)}\right]$$

we suspect that $q_\infty^{(\xi)}$ and $q_\infty^{(u)}$ may, at least in some circumstances, act as stochastic lower and upper bounds respectively to the stationary queue size q_∞ induced by $M|G|\infty$ arrivals. As a first step in this direction we now establish a comparison between $q_\infty^{(\xi)}$ and $q_\infty^{(u)}$, in the increasing convex stochastic ordering sense.

Proposition 2.3.1 *If* $\lambda \mathbf{E}[\sigma] < c$ *and* $\mathbf{E}\left[\sigma^2\right] < \infty$, *then the stationary queue lengths* $q_\infty^{(\xi)}$ *and* $q_\infty^{(u)}$ *associated with recursions (2.52) and (2.17) satisfy*

$$q_\infty^{(\xi)} \leq_{icx} q_\infty^{(u)}. \tag{2.55}$$

The proof of Proposition 2.3.1 relies on the following fact:

Lemma 2.3.1 *For Poisson rvs* $X_{\gamma\lambda}$ *and* X_λ *with parameters* $\gamma\lambda$ *and* λ, *respectively, it holds that*

$$X_{\gamma\lambda} \leq_{cx} \gamma X_\lambda, \qquad \gamma \geq 1. \tag{2.56}$$

Proof. We first prove a similar comparison result for certain Bernoulli rvs; these are subsequently used to construct the Poisson rvs of interest. Let $W^{(p)}$ denote a generic Bernoulli rv with parameter p, $0 \leq p \leq 1$, i.e.,

$$\mathbf{P}\left[W^{(p)} = 1\right] = p = 1 - \mathbf{P}\left[W^{(p)} = 0\right].$$

Fix some integer $n > \gamma\lambda$ and consider two sequences of i.i.d. Bernoulli rvs $\left\{W_i^{(\frac{\lambda}{n})},\ i = 1, 2, \ldots, n\right\}$ and $\left\{W_i^{(\frac{\lambda\gamma}{n})},\ i = 1, 2, \ldots, n\right\}$ with generic rvs $W^{(\frac{\lambda}{n})}$ and $W^{(\frac{\lambda\gamma}{n})}$, respectively. For all convex mappings $\phi : \mathbb{R} \to \mathbb{R}$ and $\gamma \geq 1$, it holds that

$$\phi(\gamma) - \phi(0) - \gamma(\phi(1) - \phi(0)) \geq 0.$$

Therefore, upon comparing

$$\mathbf{E}\left[\phi\left(W^{(\frac{\lambda\gamma}{n})}\right)\right] = \frac{\gamma\lambda}{n}\,\phi(1) + \left(1 - \frac{\gamma\lambda}{n}\right)\phi(0)$$

with

$$\mathbf{E}\left[\phi\left(\gamma W^{(\frac{\lambda}{n})}\right)\right] = \frac{\lambda}{n}\,\phi(\gamma) + \left(1 - \frac{\lambda}{n}\right)\phi(0),$$

we get

$$\mathbf{E}\left[\phi\left(W^{(\frac{\lambda\gamma}{n})}\right)\right] \leq \mathbf{E}\left[\phi\left(\gamma W^{(\frac{\lambda}{n})}\right)\right].$$

Thus, by Definition B.4 (see Appendix B), we arrive at

$$W^{(\frac{\lambda\gamma}{n})} \leq_{cx} \gamma W^{(\frac{\lambda}{n})} \tag{2.57}$$

which in turn [59, Proposition 1.1.2] implies that

$$\sum_{i=1}^{n} W_i^{(\frac{\lambda\gamma}{n})} \leq_{cx} \sum_{i=1}^{n} \gamma W_i^{(\frac{\lambda}{n})}. \tag{2.58}$$

By Poisson's Convergence Theorem the rvs $X_{\lambda\gamma}$ and γX_λ can be obtained as the weak limits of the sums $\sum_{i=1}^{n} W_i^{(\frac{\lambda\gamma}{n})}$ and $\sum_{i=1}^{n} \gamma W_i^{(\frac{\lambda}{n})}$, respectively, by letting n go to infinity. The expectations of both left and right hand side of (2.58) are finite and equal to $\gamma\lambda$, the common value of the expectations of the limiting rvs. Therefore Proposition B.8 applies and (2.56) follows by taking the limit in (2.58). ■

Proof of Proposition 2.3.1. Recall that the generic rvs β and ξ are Poisson rvs with parameters λ and $\lambda\mathbf{E}[\sigma]$, respectively, with $\mathbf{E}[\sigma] \geq 1$. Applying Lemma 2.3.1, we obtain

$$\xi \leq_{cx} \beta\ \mathbf{E}[\sigma] \tag{2.59}$$

or, equivalently, by Definition B.4,

$$\mathbf{E}[\phi(\xi)] \leq \mathbf{E}[\phi(\beta\ \mathbf{E}[\sigma])] \tag{2.60}$$

for all convex mappings $\phi : \mathbb{R} \to \mathbb{R}$ for which the expectations exist. On the other hand, Jensen's inequality implies

$$\phi\left(\mathbf{E}\left[\sum_{k=1}^{\beta} \sigma_k | \beta\right]\right) \leq \mathbf{E}\left[\phi\left(\sum_{k=1}^{\beta} \sigma_k\right) | \beta\right],$$

whence

$$\mathbf{E}\left[\phi(\beta\ \mathbf{E}\left[\sigma\right])\right] \leq \mathbf{E}\left[\phi(u)\right]. \tag{2.61}$$

Combining (2.61) and (2.60) we collect $\mathbf{E}\left[\phi(\xi)\right] \leq \mathbf{E}\left[\phi(u)\right]$, which (again by Definition B.4) is tantamount to

$$\xi \leq_{cx} u. \tag{2.62}$$

Finally, we use (2.62), together with the fact that $\mathbf{E}\left[q_\infty^{(\xi)}\right]$ and $\mathbf{E}\left[q_\infty^{(u)}\right]$ exist and are both finite when $\mathbf{E}\left[\sigma^2\right] < \infty$, to conclude that (2.55) holds true by appealing to the external monotonicity of $GI|GI|1$ recursions given by Proposition B.8. ■

In the case $c = 1$ the rv $q_\infty^{(u)}$ (and $q_\infty^{(\xi)}$ as well) admits the equivalent representation (2.38) given in Section 2.3.3. This enables us to sharpen the result of Proposition 2.3.1 as follows:

Proposition 2.3.2 *Let $c = 1$ in the recursions (2.17) and (2.52). If $\lambda\mathbf{E}\left[\sigma\right] < 1$, then the stationary queue lengths $q_\infty^{(u)}$ and $q_\infty^{(\xi)}$ satisfy*

$$q_\infty^{(\xi)} \leq_{st} q_\infty^{(u)}. \tag{2.63}$$

Proof. When $c = 1$ relation (2.38) for $q_\infty^{(u)}$ is in effect, and a corresponding expression holds true for $q_\infty^{(\xi)}$. That is,

$$q_\infty^{(\xi)} =_{st} \sum_{n=1}^{\mu} (\widehat{\xi}_n - 1), \tag{2.64}$$

where the generic rv $\widehat{\xi}$ for the sequence of i.i.d. rvs $\{\widehat{\xi}_n,\ n = 1, 2, \ldots\}$ is distributed according to

$$\mathbf{P}\left[\widehat{\xi} = 0\right] = 0; \quad \mathbf{P}\left[\widehat{\xi} = n\right] = \frac{1}{\mathbf{E}\left[\xi\right]}\ \mathbf{P}\left[\xi \geq n\right], \quad n = 1, 2, \ldots, \tag{2.65}$$

and the rv μ is independent of $\{\widehat{\xi}_n,\ n = 1, 2, \ldots\}$, with

$$\mathbf{P}\left[\mu = n\right] = \mathbf{E}\left[\xi\right]^n \left(1 - \mathbf{E}\left[\xi\right]\right), \quad n = 0, 1, \ldots. \tag{2.66}$$

Noting that $\mathbf{E}\left[\xi\right] = \mathbf{E}\left[u\right]$ we have $\mu =_{st} \nu$, where the distribution of ν was given in (2.39) . Also, the convex stochastic comparison (2.62) implies

$$\mathbf{E}\left[(\xi - n)^+\right] \leq \mathbf{E}\left[(u - n)^+\right], \quad n = 0, 1, \ldots. \tag{2.67}$$

From (2.34) and (2.65) we see that

$$\mathbf{P}\left[\widehat{\xi} > n\right] = \frac{1}{\mathbf{E}\left[\xi\right]}\, \mathbf{E}\left[(\xi - n)^+\right] \quad n = 0, 1, \ldots$$

with the corresponding relation for u. Inequality (2.67) now yields

$$\mathbf{P}\left[\widehat{\xi} > n\right] \leq \mathbf{P}\left[\widehat{u} > n\right], \quad n = 0, 1, \ldots,$$

or, equivalently,

$$\widehat{\xi} \leq_{st} \widehat{u}. \tag{2.68}$$

Thus, the convex stochastic comparison between ξ and u translates into a strong stochastic comparison between $\widehat{\xi}$ and $\widehat{u}$, and the conclusion (2.63) follows from $\mu =_{st} \nu$, (2.38), (2.64), (2.68) and Proposition 2.2.5 in [59, p. 45]. ■

Chapter 3

Heavy traffic: Lévy motion limits

3.1 Introduction

In this chapter we derive the non-degenerate limiting distribution, as the traffic intensity $\lambda \mathbf{E}[\sigma]$ tends to the multiplexer release rate c, of the appropriately normalized queue length induced by an $M|G|\infty$ arrival process, for a generally distributed activity rv σ. The arising limits are classified in terms of the short– vs. long–range dependent property of the $M|G|\infty$ process, as determined by the tail behavior of σ. In the short–range dependent regime the limiting distribution distribution is exponential, as is the case in the classical $GI|G|1$ queue, originally studied by Kingman in [31, 32]. However, under long–range dependence the results do not involve the fractional Brownian motion model of Norros [43, 44]. Different self–similar limits arise in the form of Lévy motion, leading to a buffer content distribution with hyperbolic decay.

The basic idea behind the proof of these results is a "convergence together" argument which allows us to identify processes with well–known heavy traffic behavior, under both short– and long–range dependence. This is accomplished chiefly by combining standard results on stable rvs and their domain of attraction [20],

with a general functional convergence result for processes with stationary independent increments due to Skorokhod [57]. We point out that, even in the short–range dependent case, convergence to Brownian motion does not appear to follow from standard results for stationary processes [5, Thm. 20.1, p. 174], as it is not obvious that the $M|G|\infty$ busy server process satisfies the required mixing property. However, as mentioned in Proposition 2.1.2, the $M|G|\infty$ busy server process is strongly positively correlated – it is an associated process. Because of this property, it is then possible under short–range dependence to develop an alternative approach similar to that used by Newman and Wright in [42] in establishing the Invariance Principle for sequences of associated random variables. This approach is not pursued here.

3.2 The heavy traffic regime

We seek to understand the behavior of the (stable) queue with the the correlated $M|G|\infty$ arrival process, under the assumption that it is almost fully utilized, i.e., $\lambda\mathbf{E}[\sigma]$, though less than the release rate c, is very close to c. This typically involves obtaining limiting expressions of properly rescaled quantities of interest, as the packet arrival rate $\lambda\mathbf{E}[\sigma]$ tends towards its critical value c. Here, the quantity of interest is the steady–state queue size q_∞. A natural setup to investigate this problem consists of embedding the discrete–time queue with release rate c driven by an $M|G|\infty$ input process (λ, σ) into a parametric family of like queueing systems, indexed by an integer parameter, say r. More precisely, for each $r = 1, 2, \ldots$ we take the r^{th} system to be a discrete–time queue with release rate c driven by an $M|G|\infty$ input process $\{b_n^r,\ n = 0, 1, \ldots\}$ characterized by the pair (λ_r, σ). The corresponding queue size sequence $\{q_n^r,\ n = 0, 1, \ldots\}$ also obeys the Lindley

recursion (2.11), and admits a representation of the form (2.15), i.e.,

$$q_0^r = 0; \quad q_n^r = s_n^r - nc - \inf\left(s_j^r - jc,\ j = 0, 1, \ldots, n\right), \quad n = 1, 2, \ldots \tag{3.1}$$

where $\{s_n^r,\ n = 1, 2, \ldots\}$ is the sequence of partial sums (2.12) associated with $\{b_n^r,\ n = 1, 2, \ldots\}$. We take $\lambda_r \mathbf{E}[\sigma] < c$ for all $r = 1, 2, \ldots$ for some fixed $c > 0$, so that

$$\lim_{r\to\infty} \lambda_r = c/\mathbf{E}[\sigma]. \tag{3.2}$$

Thus, each one of these systems is stable with $\lim_{r\to\infty} \lambda_r \mathbf{E}[\sigma] = c$, thereby capturing the notion that "the system is driven to heavy traffic." We seek a scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ such that the convergence in distribution

$$\frac{q_\infty^r}{\zeta_r} \Longrightarrow_r Q \tag{3.3}$$

takes place to some $\mathbb{R}$–valued rv Q.

Unfortunately, this heavy traffic program cannot be carried out in this form as exact expressions are unavailable for the distribution of q_∞^r owing to the correlations present in the $M|G|\infty$ input process, and we need to resort to the following indirect approach where the buffer content is rescaled in both the time and state space variables: For each $r = 1, 2, \ldots$, we define the $\mathbb{R}$–valued continuous–time processes $\{S^r(t),\ t \geq 0\}$ and $\{Q^r(t),\ t \geq 0\}$ by

$$S^r(t) := \frac{1}{\zeta_r}\left(s_{[rt]}^r - \mathbf{E}\left[s_{[rt]}^r\right]\right) \quad \text{and} \quad Q^r(t) := \frac{q_{[rt]}^r}{\zeta_r}, \quad t \geq 0,$$

and the function $\gamma^r : \mathbb{R}_+ \to \mathbb{R}$ by

$$\gamma^r(t) := \frac{1}{\zeta_r}\left([rt]c - \mathbf{E}\left[s_{[tr]}^r\right]\right) = \frac{[rt]}{\zeta_r}\left(c - \lambda_r \mathbf{E}[\sigma]\right), \quad t \geq 0.$$

The convergence (3.3) can be stated informally as

$$\lim_{r\to\infty}\lim_{t\to\infty} Q^r(t) = Q \tag{3.4}$$

with limits understood in the sense of weak convergence. The approach to heavy traffic followed here is to interchange the order of these limits, i.e., to evaluate

$$\lim_{t\to\infty}\lim_{r\to\infty} Q^r(t) \tag{3.5}$$

which corresponds to first taking r to infinity, and then letting t go to infinity. Assuming that the limits can be taken in that order, we are then left with the task of showing that

$$\lim_{r\to\infty}\lim_{t\to\infty} Q^r(t) = Q = \lim_{t\to\infty}\lim_{r\to\infty} Q^r(t). \tag{3.6}$$

In this chapter we concentrate only on establishing the first step (3.5), and it is well known [27, 65] that the theory of weak convergence on function spaces provides a natural framework for doing so. To that end, we pause briefly to introduce the needed notation, as well as to highlight several points from the theory of weak convergence of processes; this material is drawn mostly from [5, pp. 150–153] to which the reader is referred for additional information:

For each $T > 0$, let $D[0,T]$ denote the space of mappings $[0,T] \to \mathbb{R}$ which are right–continuous with left limits; the space $D[0,T]$ can be equipped with either the uniform topology or the standard Skorokhod topology [5, p. 111]. As in [5, p. 150], a concept prefixed with U (resp. S) refers to the uniform (resp. Skorokhod) topology. For probability measures defined on the collection of U–Borel (resp. S–Borel) sets on $D[0,T]$, we refer to weak convergence in the sense of the uniform (resp. Skorokhod) topology by U–weak (resp. S–weak) convergence, and we write $\overset{U}{\Longrightarrow}_r$ (resp. $\overset{S}{\Longrightarrow}_r$) (with the understanding that r goes to infinity). For probability measures defined on the collection of U–Borel sets, U–weak convergence implies S–weak convergence but the converse is false. This implication will be used repeatedly in various technical arguments [Sections 3.5 and 3.7].

Finally, let $D[0,\infty)$ denote the space of mappings $\mathbb{R}_+ \to \mathbb{R}$ which are right–continuous with left limits. In this chapter, we present results on the S–weak convergence of the restrictions to finite intervals of sequences of $\mathbb{R}$–valued processes with sample paths in $D[0,\infty)$. More precisely, consider the sequence of $\mathbb{R}$–valued processes $\{X_r(t),\ t \geq 0\}$, $r = 1, 2, \ldots$, with sample paths in $D[0,\infty)$. Whenever for each $T > 0$ we have the S–weak convergence

$$\{X_r(t),\ 0 \leq t \leq T\} \stackrel{S}{\Longrightarrow}_r \{X(t),\ 0 \leq t \leq T\} \quad \text{in } D[0,T]$$

for some $\mathbb{R}$–valued process $\{X(t),\ t \geq 0\}$ with sample paths in $D[0,\infty)$, we simplify the notation by writing

$$\{X_r(t),\ t \geq 0\} \Longrightarrow_r \{X(t),\ t \geq 0\}.$$

Now, noting that (3.1) can be rewritten as

$$Q^r(t) = S^r(t) - \gamma^r(t) - \inf_{0 \leq x \leq t} \left(S^r(x) - \gamma^r(x)\right), \quad t \geq 0, \tag{3.7}$$

and recalling the continuous mapping theorem [5, Thm. 5.1, p. 30], we conclude that the first limit in (3.5) requires at the very least identifying a scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ that ensures the convergence

$$\{S^r(t),\ t \geq 0\} \Longrightarrow_r \{S(t),\ t \geq 0\} \tag{3.8}$$

for some non–trivial limiting process $\{S(t),\ t \geq 0\}$.

3.3 The main heavy traffic results

As will become apparent shortly, the choice of the scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ and the characterization of the limiting process $\{S(t),\ t \geq 0\}$ entering

(3.8) both depend on the distribution of the rv σ which controls the correlations in the input packet stream. It is nevertheless easy to see that in order to avoid collecting only a law of large numbers, any candidate scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ should obey the following necessary condition:

Condition (A) *The scaling sequence* $\{\zeta_r,\ r = 1, 2, \ldots\}$ *satisfies*

$$\lim_{r\to\infty} \zeta_r = +\infty \quad \textit{with} \quad \lim_{r\to\infty} \frac{\zeta_r}{r} = 0.$$

The heavy traffic assumption below refines (3.2), and guarantees that, as r goes to infinity, the family of queueing systems described by (3.7) gradually approaches instability at the appropriate speed:

Assumption (B) *The scaling sequence* $\{\zeta_r,\ r = 1, 2, \ldots\}$ *satisfies*

$$\lim_{r\to\infty} (\lambda_r \mathbf{E}[\sigma] - c)\ \frac{r}{\zeta_r} = -\gamma \quad \textit{or equivalently,} \quad \lambda_r \mathbf{E}[\sigma] = c - \frac{\zeta_r}{r}(\gamma + o(1))$$

for some $\gamma > 0$.

Condition (A) and Assumption (B) are enforced throughout. It is worth pointing out that the scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ is essentially unique, i.e., any other scaling sequence $\{\zeta'_r,\ r = 1, 2, \ldots\}$ yielding a non–degenerate limit in (3.8) must satisfy $\lim_{r\to\infty} \zeta'_r/\zeta_r = C$ for some finite constant $C > 0$.

We begin with the case where the $M|G|\infty$ process is short–range dependent and let $\{B(t),\ t \geq 0\}$ denote a standard Brownian motion.

Theorem 3.3.1 (Short–range dependence) *If* $\mathbf{E}\left[\sigma^2\right] < \infty$, *then with* $\zeta_r = \sqrt{r}$, $r = 1, 2, \ldots$, *it holds that*

$$\{S^r(t),\ t \geq 0\} \Longrightarrow_r \{\sqrt{\frac{c\mathbf{E}[\sigma^2]}{\mathbf{E}[\sigma]}}\ B(t),\ t \geq 0\}.$$

The remaining results are obtained under the additional assumption that the tail of σ is regularly varying of order α $(1 < \alpha \leq 2)$, i.e., of the form

$$\mathbf{P}\left[\sigma > n\right] = n^{-\alpha} h(n), \quad n = 1, 2, \ldots \tag{3.9}$$

for some slowly varying function $h : \mathbb{R}_+ \to \mathbb{R}_+$, in which case the mean $\mathbf{E}\left[\sigma\right]$ is finite. Of particular interest for the forthcoming discussion is the realization that the truncated second moment of σ is $(2-\alpha)$–regularly varying. Writing

$$l_\alpha(x) := \begin{cases} \dfrac{\alpha}{2-\alpha}\, h(x) & \text{if } 1 < \alpha < 2 \\ \\ 2\displaystyle\sum_{r=1}^{[x]} \frac{h(r)}{r} & \text{if } \alpha = 2 \end{cases} \tag{3.10}$$

for all $x > 0$, we can show via Proposition 3.8.1 that the function $l_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ is slowly varying and that whenever $\mathbf{E}\left[\sigma^2\right] = \infty$, it holds

$$\mathbf{E}\left[\mathbf{1}\left[\sigma \leq n\right]\sigma^2\right] \sim n^{2-\alpha} l_\alpha(n) \quad (n \to \infty). \tag{3.11}$$

The details of the proof of this asymptotic equivalence are identical to those of (3.43) and (3.44).

The next proposition handles the boundary value $\alpha = 2$, which represents a hybrid case between short– and long–range dependence.

Theorem 3.3.2 *Assume $\alpha = 2$ in (3.9) with $\mathbf{E}\left[\sigma^2\right] = \infty$. Then, with $\{\zeta_r,\ r = 1, 2, \ldots\}$ satisfying*

$$\lim_{r\to\infty} \frac{r}{\zeta_r^2} l_2(\zeta_r) = \lim_{r\to\infty} \frac{r}{\zeta_r^2} \mathbf{E}\left[\mathbf{1}\left[\sigma \leq \zeta_r\right]\sigma^2\right] = K \tag{3.12}$$

for some positive constant K, it holds that

$$\{S^r(t),\ t \geq 0\} \Longrightarrow_r \{\sqrt{\frac{cK}{\mathbf{E}\left[\sigma\right]}}\ B(t),\ t \geq 0\}.$$

Finally, we turn to the case of bona fide long–range dependence, i.e., $1 < \alpha < 2$. We recall Definition 1.3.7 and let $\{L_\alpha(t),\ t \geq 0\}$ denote a spectrally positive, α–stable Lévy motion, i.e., an α–stable Lévy motion such that for all $t > 0$, the rv $L_\alpha(t)$ is a stable rv $S_\alpha(t^{1/\alpha}, 1, 0)$ [53, p. 9] characterized by

$$\mathbf{E}\left[\exp(i\theta L_\alpha(t))\right] = \exp\left(-t|\theta|^\alpha\left(1 - i\,\mathrm{sgn}(\theta)\tan(\frac{\pi\alpha}{2})\right)\right), \quad \theta \in \mathbb{R}\,. \tag{3.13}$$

Theorem 3.3.3 (Long–range dependence) *If $1 < \alpha < 2$ in (3.9), then with $\{\zeta_r, r = 1, 2, \ldots\}$ satisfying*

$$\lim_{r\to\infty} \frac{r}{\zeta_r^\alpha} h(\zeta_r) = \lim_{r\to\infty} r\mathbf{P}\left[\sigma > \zeta_r\right] = K \tag{3.14}$$

for some positive constant K, it holds that

$$\{S^r(t),\ t \geq 0\} \Longrightarrow_r \{\ \left(C_K \cos(\pi\frac{2-\alpha}{2})\right)^{1/\alpha} L_\alpha(t),\ t \geq 0\ \} \tag{3.15}$$

where

$$C_K := \frac{cK\Gamma(2-\alpha)}{(\alpha-1)\mathbf{E}\left[\sigma\right]}. \tag{3.16}$$

We close with a characterization of the scaling sequences encountered in Theorems 3.3.2 and 3.3.3; its proof is given in Proposition A.3 of Appendix A.

Proposition 3.3.1 *The scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ of Theorems 3.3.2 and 3.3.3 is $1/\alpha$–regularly varying, $1 < \alpha \leq 2$, i.e., of the form $\zeta_r = r^{1/\alpha}\widehat{h}(r)$ for some slowly varying function $\widehat{h} : \mathbb{R}_+ \to \mathbb{R}_+$.*

3.4 Consequences and comments

Several interesting inferences follow from the heavy traffic results obtained so far.

3.4.1 Queue size

We start with the heavy traffic behavior of the normalized queue length. Whenever the convergence (3.8) holds, we can immediately conclude from (3.7) and from the continuity of the reflection mapping (via the continuous mapping theorem [5, Thm. 5.1, p. 30]) that

$$\{Q^r(t),\ t \geq 0\} \Longrightarrow_r \{Q(t),\ t \geq 0\} \tag{3.17}$$

with

$$Q(t) := S(t) - \gamma t - \inf_{0 \leq x \leq t} \left(S(x) - \gamma x\right), \quad t \geq 0. \tag{3.18}$$

The form of the limit derives from (3.7) and the fact that $\lim_{r \to \infty} \gamma^r(t) = -\gamma t$ under Assumption (B).

This observation can now be used to provide a characterization of $Q(\infty)$, the steady–state buffer content in heavy traffic, under the assumptions of Theorems 3.3.1–3.3.3.

In the short–range dependent case, Theorem 3.3.1 combines with a classical result on the reflection functional of Brownian motion [27, p. 15] to yield the following.

Theorem 3.4.1 *Under the assumptions of Theorem 3.3.1, the resulting stationary heavy–traffic buffer content is exponentially distributed, with*

$$\mathbf{P}\left[Q(\infty) > x\right] = \exp\left(-\frac{2\gamma \mathbf{E}\left[\sigma\right]}{c\mathbf{E}\left[\sigma^2\right]}\ x\right), \quad x \geq 0.$$

Theorem 3.3.2 leads via (3.17)–(3.18) to a similar result.

Theorem 3.4.2 *Under the assumptions of Theorem 3.3.2, the resulting stationary heavy–traffic buffer content is exponentially distributed, with*

$$\mathbf{P}\left[Q(\infty) > x\right] = \exp\left(-\frac{2\gamma \mathbf{E}\left[\sigma\right]}{cK}\ x\right), \quad x \geq 0.$$

Finally, in the stable case, we need to introduce the class of Mittag–Leffler functions [17, p. 206]: For each $\nu > 0$ the Mittag–Leffler function $E_\nu : \mathbb{R} \to \mathbb{R}$ is given by

$$E_\nu(x) := \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(\nu n + 1)}, \quad x \in \mathbb{R}\,. \tag{3.19}$$

Theorem 3.3.3 can be combined with established facts on the reflection functional of a Lévy process [6] to yield the following conclusions.

Theorem 3.4.3 *Under the assumptions of Theorem 3.3.3, the distribution of the resulting stationary heavy–traffic buffer content is given by*

$$\mathbf{P}\left[Q(\infty) > x\right] = E_{\alpha-1}\left(-\frac{\gamma}{C_K}\, x^{\alpha-1}\right), \quad x \geq 0, \tag{3.20}$$

and the associated heavy–traffic buffer asymptotics are hyperbolic, with

$$\mathbf{P}\left[Q(\infty) > x\right] \sim \frac{cK}{\gamma(\alpha-1)\mathbf{E}\left[\sigma\right]}\, x^{1-\alpha} \quad (x \to \infty). \tag{3.21}$$

Proof. Combining Proposition 5a of [6, p. 725] (or Theorem in [26, p. 417]) with (3.18) and Theorem 3.3.3, we obtain

$$\mathbf{E}\left[e^{-sQ(\infty)}\right] = \frac{\gamma}{\gamma + C_K s^{\alpha-1}}\,, \quad s \geq 0. \tag{3.22}$$

Application of Fubini's theorem in (3.22) yields

$$\begin{aligned} \int_0^\infty e^{-sx}\, \mathbf{P}\left[Q(\infty) > x\right]\, dx &= \frac{1}{s}\left(1 - \mathbf{E}\left[e^{-sQ(\infty)}\right]\right) \\ &= \frac{1}{s}\, \frac{C_K}{C_K + \gamma s^{1-\alpha}}\,, \quad s \geq 0 \end{aligned} \tag{3.23}$$

and a simple change of variable gives

$$\int_0^\infty e^{-x}\, \mathbf{P}\left[Q(\infty) > \frac{x}{s}\right]\, dx = \frac{C_K}{C_K + \gamma s^{1-\alpha}}\,, \quad s \geq 0. \tag{3.24}$$

Letting $s \to \infty$ we note that the mapping $x \to \mathbf{P}\left[Q(\infty) > \frac{x}{s}\right]$ monotonically increases to the constant mapping $x \to \mathbf{P}\left[Q(\infty) > 0\right]$, whence

$$\begin{aligned}
\mathbf{P}\left[Q(\infty) > 0\right] &= \int_0^\infty e^{-x} \lim_{s\to\infty} \mathbf{P}\left[Q(\infty) > \frac{x}{s}\right] \, dx \\
&= \lim_{s\to\infty} \int_0^\infty e^{-x} \, \mathbf{P}\left[Q(\infty) > \frac{x}{s}\right] \, dx \\
&= \lim_{s\to\infty} \frac{C_K}{C_K + \gamma s^{1-\alpha}} \\
&= 1
\end{aligned}$$

by the monotone convergence theorem, so that $Q(\infty)$ has no point mass at 0. For $|s| > (C_K/\gamma)^{1/(\alpha-1)}$, the right–hand side in (3.23) can be represented by an absolutely convergent geometric series, so that

$$\int_0^\infty e^{-sx} \, \mathbf{P}\left[Q(\infty) > x\right] \, dx = \frac{1}{s} \sum_{n=0}^\infty \left(-\frac{\gamma}{C_K}\right)^n s^{(1-\alpha)n}, \quad |s| > (C_K/\gamma)^{1/(\alpha-1)}.$$

Therefore, by Theorem 35.2 in [13, p. 192], the transform can be inverted term by term to yield

$$\mathbf{P}\left[Q(\infty) > x\right] = \sum_{n=0}^\infty \left(-\frac{\gamma}{C_K}\right)^n \frac{x^{(\alpha-1)n}}{\Gamma((\alpha-1)n+1)}, \quad x \geq 0$$

and (3.20) readily follows from the definition (3.19). The asymptotics (3.21) are verified by observing that

$$1 - \mathbf{E}\left[e^{-sQ(\infty)}\right] \sim \frac{C_K}{\gamma} \, s^{\alpha-1} \quad (s \to 0+)$$

and by making use of a standard Tauberian result [7, Corollary 8.1.7]. ■

3.4.2 On selecting the heavy traffic scaling

As the appropriate scaling sequence $\{\zeta_r, \ r = 1, 2, \ldots\}$ is revealing of the nature of the limiting heavy traffic process $\{S(t), \ t \geq 0\}$, we briefly discuss here its selection.

In Section 2.1.2 we mentioned that, under (3.9) with $1 < \alpha < 2$, the $M|G|\infty$ busy server process possesses the second order asymptotic self–similarity property, with parameter $(3-\alpha)/2$, i.e., by aggregating the original process $\{b_n,\ n = 0, 1, \ldots\}$ in blocks of size m and dividing by the block size, we obtain in the limit (as m goes to infinity) the same correlation function as that of a fractional Gaussian noise process. Such convergence of the correlation function tempts one to think that the appropriate scaling ensuring (3.8) might be the one that balances the rate of growth of the partial sums variance, so that convergence (3.8) occurs to a limiting process with finite variance. We now explore this point in some detail:

By standard calculations, we find the variance of the partial sums to be

$$\mathrm{var}\left[S^r(t)\right] = \frac{\lambda_r \mathbf{E}\left[\sigma\right]}{\zeta_r^2}\left([rt] + 2\sum_{k=1}^{[rt]}([rt]-k)\mathbf{P}\left[\widehat{\sigma} > k\right]\right), \quad t \geq 0$$

for all $r = 1, 2, \ldots$. It can be shown that when the tail of σ satisfies (3.9) with $1 < \alpha < 2$, the candidate scaling $\{\zeta_r,\ r = 1, 2, \ldots\}$ given by

$$\zeta_r^2 := r\sum_{k=1}^{r}\mathbf{P}\left[\widehat{\sigma} > k\right], \quad r = 1, 2, \ldots \tag{3.25}$$

indeed results in a finite limiting variance, i.e., $\lim_{r\to\infty}\mathrm{var}\left[S^r(t)\right]$ exists and is finite for all $t \geq 0$. In addition, invoking (3.94) we see that the scaling (3.25) has the asymptotic form

$$\zeta_r^2 \sim \frac{1}{(2-\alpha)(\alpha-1)\mathbf{E}\left[\sigma\right]}\, r^{3-\alpha}h(r) \quad (r \to \infty). \tag{3.26}$$

and is therefore regularly varying of order $(3-\alpha)/2$.

On the other hand, from Theorem 1.3.1 we already know that convergence of a normalized partial sum process, such as $\{S^r(t),\ t \geq 0\}$, can only be to a self–similar process, and that the corresponding Hurst parameter H may be determined

through the regularly varying scaling $\{\zeta_r,\ r = 1, 2, \ldots\}$ by

$$\lim_{r\to\infty} \frac{\zeta_{[rx]}}{\zeta_r} = x^H, \quad x > 0.$$

Thus, the candidate scaling (3.25), which balances the growth of the variance, suggests possible convergence to a fractional Brownian motion with Hurst parameter $(3 - \alpha)/2$.

In the present heavy traffic setup however, convergence of the rescaled $M|G|\infty$ process to a fractional Brownian motion does not take place. The candidate scaling (3.25) is not the appropriate scaling; it is too strong and yields convergence to a degenerate limit – the identically zero process. Theorem 3.3.3, in conjuction with Proposition 3.3.1, clearly shows that the correct scaling does not contain any $r^{(3-\alpha)/2}$ factor, but instead contains the weaker $r^{1/\alpha}$ factor associated with the stable law to which the service rv σ is attracted. As a result, the limiting heavy traffic process turns out to be not a fractional Brownian motion but an α–stable $1/\alpha$–self–similar Lévy motion, the stable analog of standard Brownian motion, which has independent increments with infinite variance. In heavy traffic, the corresponding queue length asymptotics are not Weibullian, but hyperbolic with power $1 - \alpha$. Thus, $M|G|\infty$ processes demonstrate that, within long–range dependence, fractional Brownian motion does not assume the ubiquitous role that standard Brownian motion plays in the short–range dependence setup, and that modeling possibilities attracted to non–Gaussian limits are not so hard to find. Clearly, the extent to which such non–Gaussian processes can serve as useful traffic models deserves further consideration.

3.5 Outline of proof and preliminary results

In this section we organize the proof of Theorems 3.3.1–3.3.3 into a series of steps which we formalize as Propositions; their proofs are given in Section 3.7.

Look at the r^{th} queueing system for some $r = 1, 2, \ldots$, and fix $n = 0, 1, \ldots$. We note the decomposition $b_n^r = b_n^{(0)r} + b_n^{(a)r}$ where the rvs $b_n^{(0)r}$ and $b_n^{(a)r}$ describe the contributions to the number of customers in the system at the beginning of slot $[n, n+1)$ from those initially present (at $n = 0$) and from the new arrivals, respectively. It is easy to see that

$$b_n^{(0)r} = \sum_{j=1}^{b^r} \mathbf{1}\left[\widehat{\sigma}_j > n\right] \quad \text{and} \quad b_n^{(a)r} = \sum_{k=1}^{n} \sum_{j=1}^{\beta_k^r} \mathbf{1}\left[\sigma_{k,j} > n-k\right].$$

It was shown in [47, Sec. 5] that

$$s_n^{(0)r} := \sum_{j=1}^{n} b_j^{r(0)} = \sum_{j=1}^{b^r} \min(n, \widehat{\sigma}_j - 1) \tag{3.27}$$

and

$$s_n^{(a)r} := \sum_{k=1}^{n} b_k^{(a)r} = \sum_{k=1}^{n} \sum_{j=1}^{\beta_k^r} \min(\sigma_{k,j}, n-k+1). \tag{3.28}$$

We introduce the rescaled versions

$$S^{(0)r}(t) := \frac{1}{\zeta_r}\left(s_{[rt]}^{(0)r} - \mathbf{E}\left[s_{[rt]}^{(0)r}\right]\right), \quad t \geq 0$$

and

$$S^{(a)r}(t) := \frac{1}{\zeta_r}\left(s_{[rt]}^{(a)r} - \mathbf{E}\left[s_{[rt]}^{(a)r}\right]\right), \quad t \geq 0$$

so that

$$S^r(t) = S^{(0)r}(t) + S^{(a)r}(t), \quad t \geq 0. \tag{3.29}$$

Also, for each $T > 0$, the identically zero mapping on $[0, T]$ is the element of $D[0, T]$ denoted by θ_T, i.e., $\theta_T : [0, T] \to \mathbb{R}$ with $\theta_T(t) = 0$, $0 \leq t \leq T$.

We first show that the initial condition plays no role in the heavy traffic limit, as should be expected. This reduction step, as well as others taken in this section, is accomplished under the following sufficient condition.

Condition (B) *The scaling sequence* $\{\zeta_r,\ r = 1, 2, \ldots\}$ *satisfies*

$$\lim_{r\to\infty} \frac{1}{\zeta_r} \sum_{j=1}^{r} \mathbf{P}\left[\widehat{\sigma} > j\right] = 0.$$

Condition (B) holds under each set of assumptions of Theorems 3.3.1–3.3.3; this is shown in Proposition 3.6.1 of Section 3.6.

Proposition 3.5.1 *Under Condition (B), for each* $T > 0$ *it holds that*

$$\{S^{(0)r}(t),\ 0 \le t \le T\} \stackrel{U}{\Longrightarrow}_r \theta_T \quad \text{in } D[0, T].$$

Thus, in order to get (3.8) it suffices to consider the limiting behavior of the rescaled process $\{S^{(a)r}(t),\ t \ge 0\}$. To that end, for each $r = 1, 2, \ldots$, we introduce the sequence $\{w_n^r,\ n = 0, 1, \ldots\}$ given by

$$w_0^r := 0, \quad w_n^r := \sum_{k=1}^{n} \sum_{j=1}^{\beta_k^r} \sigma_{k,j}, \quad n = 1, 2, \ldots \tag{3.30}$$

which can be interpreted as the sequence of partial sums associated with the instantaneous arrivals (2.16). The corresponding rescaled process $\{W^r(t),\ t \ge 0\}$ is given by

$$W^r(t) := \frac{1}{\zeta_r} \left(w_{[rt]}^r - \mathbf{E}\left[w_{[rt]}^r\right]\right), \quad t \ge 0. \tag{3.31}$$

The main idea driving the discussion is that in as much as heavy traffic is concerned, the process $\{W^r(t),\ t \ge 0\}$ acts as a surrogate for $\{S^{(a)r}(t),\ t \ge 0\}$. This is made precise through the following "convergence together" result.

Proposition 3.5.2 *Under Condition (B), for each $T > 0$ it holds that*

$$\{W^r(t) - S^{(a)r}(t),\ 0 \le t \le T\} \stackrel{U}{\Longrightarrow}_r \theta_T \quad \text{in } D[0,T].$$

Combining Propositions 3.5.1 and 3.5.2, we immediately get the following conclusion from the decomposition (3.29).

Corollary 3.5.1 *Under Condition (B), for each $T > 0$ it holds that*

$$\{W^r(t) - S^r(t),\ 0 \le t \le T\} \stackrel{U}{\Longrightarrow}_r \theta_T \quad \text{in } D[0,T],$$

so that the process $\{S^r(t),\ 0 \le t \le T\}$ is S–weakly convergent if and only if $\{W^r(t),\ 0 \le t \le T\}$ is S–weakly convergent, and convergence is to the same limit.

Thus, we need only consider the convergence of the process $\{W^r(t),\ t \ge 0\}$, and characterize the limiting process. In fact, a further reduction can be achieved by noting that in heavy traffic we can replace $\{\beta_k^r,\ k = 1, 2, \ldots\}$ by the limiting i.i.d. sequence $\{\beta_k,\ k = 1, 2, \ldots\}$, where the generic rv β is a Poisson rv with parameter $c/\mathbf{E}[\sigma]$. More precisely, consider the modified workload process $\{v_n,\ n = 0, 1, \ldots\}$ given by

$$v_0 = 0; \quad v_n = \sum_{k=1}^{n} \sum_{j=1}^{\beta_k} \sigma_{k,j}, \quad n = 1, 2, \ldots \tag{3.32}$$

under the assumption that the rvs $\{\beta_k,\ k = 1, 2, \ldots\}$ are independent of the session duration rvs $\{\sigma_{n,j},\ n, j = 1, 2, \ldots\}$. For each $r = 1, 2, \ldots$, the corresponding rescaled process $\{V^r(t),\ t \ge 0\}$ is defined by

$$V^r(t) := \frac{1}{\zeta_r} \left(v_{[rt]} - \mathbf{E}\left[v_{[rt]}\right] \right), \quad t \ge 0. \tag{3.33}$$

Proposition 3.5.3 *Under Assumption (B), the process $\{W^r(t),\ 0 \le t \le T\}$ is S–weakly convergent if and only if $\{V^r(t),\ 0 \le t \le T\}$ is S–weakly convergent, and convergence is to the same limit.*

Corollary 3.5.1 and Proposition 3.5.3 together lead to the following conclusion:

Corollary 3.5.2 *Under Assumption (B) and Condition (B), the process* $\{S^r(t),\ 0 \le t \le T\}$ *is* S*–weakly convergent if and only if* $\{V^r(t),\ 0 \le t \le T\}$ *is* S*–weakly convergent, and convergence is to the same limit.*

3.6 Proofs of Theorems 3.3.1–3.3.3

First, the big picture: Corollary 3.5.2 and Proposition 3.6.1 (given below) imply that in proving Theorems 3.3.1–3.3.3 we need only investigate the convergence of the modified workload process (3.33). This is a much easier task as we now deal with the (normalized) partial sums process associated with a single sequence of i.i.d. rvs, of finite mean but possibly infinite variance, an extensively studied situation where the (functional form of the) classical Central Limit Theorem and its generalization to i.i.d. summands with infinite variance, are expected to yield the requested convergence. In fact, as we shall see shortly, the convergence of the finite dimensional distributions of $\{V^r(t),\ t \ge 0\}$ turns out to be an easy by–product of classical results concerning stable distributions and their domains of attraction [20, pp. 574–581]. Finally, the desired S–weak convergence of the process $\{V^r(t),\ t \ge 0\}$, thus of $\{S^r(t),\ t \ge 0\}$, will be validated through functional convergence results due to Skorokhod [57]. This approach clearly explains the form of the results obtained in this chapter, providing insights as to when the process $\{V^r(t),\ t \ge 0\}$ is expected to converge, and to which limit. A different, analytic approach using characteristic functions was pursued in the technical report [60].

We now proceed with the details: In Section 3.8 we give a proof that the technical Condition (B) required to establish the "convergence together" argument,

indeed holds under the assumptions of Theorems 3.3.1–3.3.3.

Proposition 3.6.1 *Condition (B) holds true for each of the scaling sequences* $\{\zeta_r,\ r = 1, 2, \ldots\}$ *in Theorems 3.3.1–3.3.3.*

Next, we consider the generic compound rv Y given by

$$Y := \sum_{j=1}^{\beta} \sigma_j \tag{3.34}$$

where the rv β is a Poisson rv with parameter $c/\mathbf{E}[\sigma]$ and independent of the i.i.d. rvs $\{\sigma_j,\ j = 1, 2, \ldots\}$ which are distributed according to σ. Fixing $t \geq 0$, we note that

$$V^r(t) =_{st} \frac{1}{\zeta_r} \sum_{k=1}^{[rt]} (Y_k - \mathbf{E}[Y_k]), \quad r = 1, 2, \ldots \tag{3.35}$$

where the i.i.d. rvs $\{Y_k,\ k = 1, 2, \ldots\}$ are distributed according to the generic rv Y.

For easy reference, we restate some useful facts concerning stable distributions and their domains of attraction; the reader is refered to [20, pp. 574–581] for additional material: Let L be a rv with distribution not concentrated at one point, and let $\{X_r,\ r = 1, 2, \ldots\}$ be a sequence of i.i.d. rvs, with generic rv X. We say that X belongs to the domain of attraction of the rv L if there exist normalizing constants $\zeta_r > 0$ and $c_r, r = 1, 2, \ldots$, such that

$$\frac{X_1 + \ldots + X_r - rc_r}{\zeta_r} \Longrightarrow_r L. \tag{3.36}$$

By Theorem 1 of [20, p. 576] only stable rvs possess a domain of attraction. By Theorem 2 in [20, p. 577], in order for X to belong to the domain of attraction of a stable law with exponent α, $0 < \alpha \leq 2$, it is necessary that its truncated second moment be regularly varying with exponent $2 - \alpha$, i.e.,

$$\mathbf{E}\left[\mathbf{1}[X \leq r] X^2\right] \sim r^{2-\alpha} g(r) \quad (r \to \infty), \tag{3.37}$$

for some slowly varying function $g : \mathbb{R}_+ \to \mathbb{R}_+$. The associated scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ in (3.36) must then satisfy

$$\lim_{r\to\infty} \frac{r}{\zeta_r^2} \mathbf{E}\left[\mathbf{1}\left[X \leq \zeta_r\right] X^2\right] = M \tag{3.38}$$

for some constant $M > 0$ [20, p. 579]. Moreover, if $\mathbf{E}[X]$ is finite, then by Theorem 3(ii) of [20, p. 581] we can take $c_r = \mathbf{E}[X]$, $r = 1, 2, \ldots$.

We are now ready to discuss Theorems 3.3.1–3.3.3 which are all proven in the same manner, although for clarity of presentation, we shall consider each of them separately. As $\mathbf{E}[Y]$ is finite under the enforced assumptions, we conclude from (3.35) and (3.36) that for each $t > 0$, the convergence question concerning $\{V^r(t),\ r = 1, 2, \ldots\}$ is equivalent to determining whether the rv Y is attracted to a stable law, and to which one. In asserting this equivalence we rely on the fact that the scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ so selected is regularly varying, as turns out to be the case by Proposition 3.3.1, so that

$$\lim_{r\to\infty} \frac{\zeta_{[rt]}}{\zeta_r} = t^{1/\alpha}, \quad t \geq 0. \tag{3.39}$$

In each case, we show that both the necessary condition (3.37) and the accompanying sufficient condition stated in [20, p. 577] are satisfied. This occurs simply because the generic rv Y inherits the tail behavior of the generic service time σ under each set of assumptions of Theorems 3.3.1–3.3.3.

A proof of Theorem 3.3.1.

Since $\mathbf{E}\left[\sigma^2\right] < \infty$, the variance of Y is also finite, and is given by

$$\mathrm{var}[Y] = \mathrm{var}[\beta]\,\mathbf{E}[\sigma]^2 + \mathbf{E}[\beta]\,\mathrm{var}[\sigma] = \frac{c\mathbf{E}[\sigma^2]}{\mathbf{E}[\sigma]}. \tag{3.40}$$

Hence, the truncated second moment of Y varies slowly, i.e., (3.37) holds with $\alpha = 2$ and as Y is never degenerate at one point, it follows from Corollary 1 to

Theorem 2 in [20, p. 578] that Y is attracted to the normal distribution. Obviously, the scaling $\zeta_r = \sqrt{r},\ r = 1, 2, \ldots$ satisfies (3.38), with $M = c\mathbf{E}\left[\sigma^2\right]/\mathbf{E}\left[\sigma\right]$. In fact, by a well–known result of Donsker [5, Thm. 16.1, p. 137], selecting $\zeta_r = \sqrt{r},\ r = 1, 2, \ldots$ ensures that the process $\{V^r(t),\ t \geq 0\}$ is S–weakly convergent to a Brownian motion, with

$$\{V^r(t),\ t \geq 0\} \stackrel{S}{\Longrightarrow}_r \{\sqrt{M}\ B(t),\ t \geq 0\}. \tag{3.41}$$

Combining (3.41) with Proposition 3.6.1 and Corollary 3.5.2 immediately concludes the proof. ■

Under the assumptions of Theorems 3.3.2 and 3.3.3, $\mathbf{E}\left[\sigma^2\right]$ is infinite, and the compound Poisson rv Y now has infinite variance. Also, if σ satisfies the tail condition (3.9), so does Y with

$$\mathbf{P}\left[Y > r\right] = \mathbf{P}\left[\sum_{j=1}^{\beta} \sigma_j > r\right] \sim \mathbf{E}\left[\beta\right]\ r^{-\alpha} h(r) \quad (r \to \infty). \tag{3.42}$$

The asymptotic equality in (3.42) is stated as an exercise in [20, Ex. 31, p. 288], where the reader will find hints for its proof (see also [16]). Next, we check that the truncated second moment of Y is given by

$$\mathbf{E}\left[\mathbf{1}\left[Y \leq r\right] Y^2\right] = 2\sum_{n=1}^{r} n\mathbf{P}\left[Y > n\right] - r(r+2)\mathbf{P}\left[Y > r\right] + \sum_{n=0}^{r-1} \mathbf{P}\left[Y > n\right]$$

for each $r = 1, 2, \ldots$. Using (3.42) in this last expression, we find that

$$\mathbf{E}\left[\mathbf{1}\left[Y \leq r\right] Y^2\right] \sim \mathbf{E}\left[\beta\right] \left(2\sum_{n=1}^{r} n^{1-\alpha} h(n) - r^{2-\alpha} h(r)\right) \quad (r \to \infty) \tag{3.43}$$

because $\mathbf{E}\left[Y\right]$ is finite and $\mathbf{E}\left[Y^2\right]$ infinite. We close these preliminary remarks by noting that the truncated second moments of σ and Y are obviously related to each other by

$$\mathbf{E}\left[\mathbf{1}\left[Y \leq r\right] Y^2\right] \sim \mathbf{E}\left[\beta\right] \mathbf{E}\left[\mathbf{1}\left[\sigma \leq r\right] \sigma^2\right] \quad (r \to \infty). \tag{3.44}$$

A proof of Theorem 3.3.2.

Inserting $\alpha = 2$ in (3.43) and using the definition (3.10) (with $\alpha = 2$), we get

$$\mathbf{E}\left[\mathbf{1}\left[Y \le r\right] Y^2\right] \sim \mathbf{E}\left[\beta\right]\ \left(l_2(r) - h(r)\right) \quad (r \to \infty).$$

By Proposition 3.8.1(ii), $l_2 : \mathbb{R}_+ \to \mathbb{R}_+$ is slowly varying with

$$\lim_{r\to\infty} \frac{h(r)}{l_2(r)} = 0,$$

so that

$$\mathbf{E}\left[\mathbf{1}\left[Y \le r\right] Y^2\right] \sim \mathbf{E}\left[\beta\right]\ l_2(r) \quad (r \to \infty). \tag{3.45}$$

This time, by Corollary 1 in [20, p. 578], the slow variation of the truncated second moment is a necessary and sufficient condition for Y to be attracted to the normal distribution, with normalizing coefficients selected according to (3.38) (despite the fact that the variance of Y is now infinite). Since the marginals of the process $\{V^r(t),\ t \ge 0\}$, which has stationary, independent increments, converge to a Gaussian distribution, it follows by [57, Theorem 2.7] without any additional conditions that (3.41) takes place. Because of (3.45), selecting the scaling $\{\zeta_r,\ r = 1, 2, \ldots\}$ according to (3.38), with $M = cK/\mathbf{E}\left[\sigma\right]$, is equivalent to (3.12). Combining (3.41) with Proposition 3.6.1 and Corollary 3.5.2 completes the proof. ■

A proof of Theorem 3.3.3.

When $1 < \alpha < 2$ in (3.9) the rvs σ and Y have infinite second moment, and Proposition 3.8.1(i) implies

$$\lim_{r\to\infty} \frac{1}{r^{2-\alpha}h(r)} \sum_{n=1}^{r} n^{1-\alpha} h(n) = \frac{1}{(2-\alpha)}. \tag{3.46}$$

Using this asymptotic in (3.43) we get

$$\mathbf{E}\left[\mathbf{1}\left[Y \le r\right] Y^2\right] \sim \frac{\alpha}{2-\alpha}\, \mathbf{E}\left[\beta\right]\; r^{2-\alpha} h(r) \quad (r \to \infty). \tag{3.47}$$

Invoking Corollary 2 of [20, p. 578], we see that (3.47) and the tail condition (3.42) are sufficient to ensure membership of Y in the domain of attraction of a non–normal stable distribution with exponent $1 < \alpha < 2$. The associated scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$, selected according to (3.38), yields convergence of the marginal distribution of $V^r(1)$, as r goes to infinity, to that of an α–stable rv, i.e.,

$$\lim_{r\to\infty} \mathbf{E}\left[\exp(i\theta V^r(1))\right] = \mathbf{E}\left[\exp\left(i\theta\left(\frac{M\Gamma(3-\alpha)}{\alpha(\alpha-1)}\cos(\pi\frac{2-\alpha}{2})\right)^{1/\alpha} L_\alpha(1)\right)\right]$$

for all θ in $\mathbb{R}$. The exact value of the constant given above can be easily verified, by recalling the expression (3.13) for the characteristic function of $L_\alpha(1)$ and comparing it with Eq. (3.18) of [20, p. 730] (note the unfortunate error in the $\pm$ sign). Next, appealing to [57, Theorem 2.7] again, we conclude that convergence of the marginals also implies S–weak convergence of the process $\{V^r(t),\ t \ge 0\}$, which has stationary, independent increments, to an α–stable Lévy motion. More precisely, it holds that

$$\{V^r(t),\ t \ge 0\} \stackrel{S}{\Longrightarrow}_r \{\left(\frac{M\Gamma(3-\alpha)}{\alpha(\alpha-1)}\cos(\pi\frac{2-\alpha}{2})\right)^{1/\alpha} L_\alpha(t),\ t \ge 0\}. \tag{3.48}$$

Using (3.47) in (3.38) with $M = cK\alpha/(2-\alpha)\mathbf{E}\left[\sigma\right]$ we obtain the scaling sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ given in (3.14). Finally, combining (3.48) with Proposition 3.6.1 and Corollary 3.5.2 shows that (3.15) holds true. ■

3.7 Proofs of Propositions 3.5.1, 3.5.2 and 3.5.3

We start by remarking that if the sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ is regularly varying (as stated in Proposition 3.3.1), then Condition (B) also implies

$$\lim_{r\to\infty} \frac{1}{\zeta_r} \sum_{j=1}^{[rt]} \mathbf{P}\left[\widehat{\sigma} > j\right] = 0, \quad t \geq 0. \tag{3.49}$$

All three proofs given in this section follow the same pattern, and are based on the following simple idea: Consider a sequence of $\mathbb{R}$–valued processes $\{X^r(t),\ t \geq 0\}$, $r = 1, 2, \ldots$, with sample paths in $D[0, \infty)$. Fix $T > 0$. According to Theorem 4.1 of [5, p. 25], the U–weak convergence

$$\{X^r(t),\ 0 \leq t \leq T\} \stackrel{U}{\Longrightarrow}_r \theta_T \quad \text{in } D[0, T],$$

follows from the convergence in probability

$$\sup_{0\leq t\leq T} |X^r(t)| \stackrel{P}{\longrightarrow}_r 0. \tag{3.50}$$

A proof of Proposition 3.5.1.

Fix $r = 1, 2, \ldots$, and note from (3.27) that

$$\sup_{0\leq t\leq T} |S^{(0)r}(t)| \leq \frac{1}{\zeta_r} \sum_{j=1}^{b^r} \min(\widehat{\sigma}_j - 1, [rT]) + \frac{\lambda_r \mathbf{E}\left[\sigma\right]}{\zeta_r}\ \mathbf{E}\left[\min(\widehat{\sigma} - 1, [rT])\right].$$

Hence, for every $\varepsilon > 0$, it is plain that

$$\begin{aligned}
&\mathbf{P}\left[\sup_{0\leq t\leq T} |S^{(0)r}(t)| > \varepsilon\right] \\
&\leq\ \mathbf{P}\left[\frac{1}{\zeta_r} \sum_{j=1}^{b^r} \min(\widehat{\sigma}_j - 1, [rT]) + \frac{\lambda_r \mathbf{E}\left[\sigma\right]}{\zeta_r}\ \mathbf{E}\left[\min(\widehat{\sigma} - 1, [rT])\right] > \varepsilon\right] \\
&\leq\ \frac{2\lambda_r \mathbf{E}\left[\sigma\right]}{\varepsilon\zeta_r}\ \mathbf{E}\left[\min(\widehat{\sigma} - 1, [rT])\right]
\end{aligned} \tag{3.51}$$

where the last step follows by Chebyshev's inequality. It is also the case that

$$\begin{aligned}\mathbf{E}\left[\min(\widehat{\sigma}-1,[rT])\right] &= \sum_{n=0}^{[rT]-1}\mathbf{P}\left[\min(\widehat{\sigma}-1,[rT])>n\right]\\ &= \sum_{n=0}^{[rT]-1}\mathbf{P}\left[\widehat{\sigma}-1>n\right]\\ &= \sum_{n=1}^{[rT]}\mathbf{P}\left[\widehat{\sigma}>n\right].\end{aligned}$$

Appealing to Condition (B) and (3.49), we get

$$\lim_{r\to\infty}\mathbf{E}\left[\frac{1}{\zeta_r}\min(\widehat{\sigma}-1,[rT])\right]=0 \tag{3.52}$$

and the conclusion

$$\lim_{r\to\infty}\mathbf{P}\left[\sup_{0\le t\le T}|S^{(0)r}(t)|>\varepsilon\right]=0$$

immediately obtains from (3.2) upon letting r go to infinity in (3.51). ■

A proof of Proposition 3.5.2.

Fix $r=1,2,\ldots$. From (3.28) and (3.30) we note that

$$\begin{aligned}w_n^r-s_n^{(a)r} &= \sum_{k=1}^{n}\sum_{j=1}^{\beta_k^r}(\sigma_{k,j}-(n-k+1))^+\\ &= \sum_{k=1}^{n}\sum_{j=1}^{\beta_{n-k+1}^r}(\sigma_{n-k+1,j}-k)^+\\ &=_{st} \sum_{k=1}^{n}\sum_{j=1}^{\beta_k^r}(\sigma_{k,j}-k)^+,\quad n=1,2,\ldots\end{aligned} \tag{3.53}$$

where the last step made use of the mutual independence of the families of i.i.d. rvs $\{\beta_k^r,\ k=1,2,\ldots\}$ and $\{\sigma_{k,j},\ k,j=1,2,\ldots\}$. It is now straightforward to check that

$$\sup_{0\le t\le T}|W^r(t)-S^{(a)r}(t)|\le_{st}\frac{1}{\zeta_r}\sum_{k=1}^{[rT]}\sum_{n=1}^{\beta_k^r}(\sigma_{k,n}-k)^+ +\frac{\lambda_r}{\zeta_r}\sum_{k=1}^{[rT]}\mathbf{E}\left[(\sigma-k)^+\right].$$

By Chebyshev's inequality, for every $\varepsilon > 0$ we obtain

$$\begin{aligned} \mathbf{P}\left[\sup_{0\le t\le T}|W^r(t) - S^{(a)r}(t)| > \varepsilon\right] &\le \frac{2\lambda_r}{\varepsilon\zeta_r}\sum_{k=1}^{[rT]}\mathbf{E}\left[(\sigma-k)^+\right] \\ &= \frac{2\lambda_r\mathbf{E}\left[\sigma\right]}{\varepsilon\zeta_r}\sum_{k=1}^{[rT]}\mathbf{P}\left[\widehat{\sigma} > k\right], \end{aligned} \tag{3.54}$$

and the desired convergence

$$\sup_{0\le t\le T}|W^r(t) - S^{(a)r}(t)| \xrightarrow{P}_r 0$$

follows upon letting r go to infinity in the upper bound (3.54), and making use of (3.2), Condition (B) and (3.49). ■

The proof of Proposition 3.5.3 requires estimates that derive from various martingales inequalities; we now state them in Lemmas 3.7.1 and 3.7.2 for easy reference: Consider a collection of integrable rvs $\{X_i,\ i = 1,\ldots,n\}$ adapted with respect to the filtration $\{\mathcal{F}_i,\ i = 1,\ldots,n\}$, i.e., for each $i = 1,\ldots,n$, the rv X_i is $\mathcal{F}_i$–measurable. We also write

$$S_i = X_1 + \ldots + X_i, \quad i = 1,\ldots,n.$$

Kolmogorov's maximal inequality [25, Corollary 2.1, p. 14] is given first.

Lemma 3.7.1 *Assume $\{(S_i,\mathcal{F}_i),\ i = 1,\ldots,n\}$ to form a martingale. Then, for each $p \ge 1$, it holds that*

$$\mathbf{P}\left[\max_{i=1,\ldots,n}|S_i| > \lambda\right] \le \lambda^{-p}\mathbf{E}\left[|S_n|^p\right], \quad \lambda > 0.$$

The von Bahr–Esseen inequality [64] is next.

Lemma 3.7.2 *Assume* $\{(X_i, \mathcal{F}_i),\ i = 1, \ldots, n\}$ *to form a martingale difference. If* $\mathbf{E}\left[|X_i|^p\right] < \infty$ *for all* $i = 1, \ldots, n$, *then*

$$\mathbf{E}\left[|S_n|^p\right] \leq 2 \sum_{i=1}^{n} \mathbf{E}\left[|X_i|^p\right], \quad 1 \leq p \leq 2.$$

In what follows Lemmas 3.7.1 and 3.7.2 are applied to the special case when the rvs $\{X_i,\ i = 1, \ldots, n\}$ are zero–mean i.i.d. rvs.

A proof of Proposition 3.5.3.

Recall that the rvs $\{\beta_k,\ k = 1, 2, \ldots\}$ are i.i.d. Poisson rvs with parameter $c/\mathbf{E}\left[\sigma\right]$, which are independent of the sequence of i.i.d. session duration rvs $\{\sigma_{k,j},\ k, j = 1, 2, \ldots\}$.

Fix $r = 1, 2, \ldots$. On the same probability triple $(\Omega, \mathcal{F}, \mathbf{P})$ where the previously mentioned rvs are defined, we introduce a family of i.i.d. $\{0, 1\}$–valued rvs $\{U^r_{k,j},\ k, j = 1, 2, \ldots\}$, i.e.,

$$\mathbf{P}\left[U^r = 1\right] = \frac{\lambda_r \mathbf{E}\left[\sigma\right]}{c} = 1 - \mathbf{P}\left[U^r = 0\right]$$

where U^r denotes the generic rv for this i.i.d. sequence. The rvs $\{U^r_{k,j},\ k, j = 1, 2, \ldots\}$ are assumed independent of the collections of rvs mentioned so far. Next, we define the rvs $\{\widetilde{\beta}^r_k, k = 1, 2, \ldots\}$ by

$$\widetilde{\beta}^r_k := \sum_{j=1}^{\beta_k} U^r_{k,j}, \quad k = 1, 2, \ldots$$

We also define the workload process $\{\widetilde{w}^r_n,\ n = 0, 1, \ldots\}$ corresponding to $\{\widetilde{\beta}^r_k,\ k = 1, 2, \ldots\}$ by

$$\widetilde{w}^r_0 := 0, \quad \widetilde{w}^r_n := \sum_{k=1}^{n} \sum_{j=1}^{\widetilde{\beta}^r_k} \sigma_{k,j}, \quad n = 1, 2, \ldots$$

and its rescaled version $\{\widetilde{W}^r(t),\ t \geq 0\}$ by

$$\widetilde{W}^r(t) := \frac{1}{\zeta_r}\left(\widetilde{w}^r_{[rt]} - \mathbf{E}\left[\widetilde{w}^r_{[rt]}\right]\right), \quad t \geq 0.$$

Under the enforced independence assumptions, it is easy to check that $\{\widetilde{\beta}^r_k,\ k = 1, 2, \ldots\} =_{st} \{\beta^r_k,\ k = 1, 2, \ldots\}$, and that

$$\{\widetilde{W}^r(t),\ 0 \leq t \leq T\} =_{st} \{W^r(t),\ 0 \leq t \leq T\}.$$

Moreover, these rvs are all defined on the same probability triple as the rescaled process $\{V^r(t),\ t \geq 0\}$. Thus, the result will be established if it holds that

$$\{V^r(t) - \widetilde{W}^r(t),\ 0 \leq t \leq T\} \overset{U}{\Longrightarrow}_r \theta_T \quad \text{in } D[0, T],$$

or equivalently, if we can show that

$$\sup_{0 \leq t \leq T} |V^r(t) - \widetilde{W}^r(t)| \xrightarrow{P}_r 0. \tag{3.55}$$

To that end, for each $r = 1, 2, \ldots$, we note from the definitions that

$$\widetilde{w}^r_n = \sum_{k=1}^{n} \sum_{j=1}^{\beta_k} U^r_{k,j}\sigma_{k,j}, \quad n = 1, 2, \ldots$$

so that

$$v^r_n - \widetilde{w}^r_n = \sum_{k=1}^{n} \sum_{j=1}^{\beta_k} (1 - U^r_{k,j})\sigma_{k,j}, \quad n = 1, 2, \ldots \tag{3.56}$$

The rvs $\{Z^r_k,\ k = 1, 2, \ldots\}$ defined by

$$Z^r_k := \sum_{j=1}^{\beta_k} (1 - U^r_{k,j})\sigma_{k,j}, \quad k = 1, 2, \ldots \tag{3.57}$$

are i.i.d., and denote by Z^r the corresponding generic rv associated with this collection of rvs. It is plain from (3.56) and (3.57) that

$$\sup_{0 \leq t \leq T} |V^r(t) - \widetilde{W}^r(t)| = \frac{1}{\zeta_r} \sup_{1 \leq n \leq [rT]} \left| \sum_{k=1}^{n} (Z^r_k - \mathbf{E}\left[Z^r_k\right]) \right|.$$

Fix $\varepsilon > 0$. Invoking the maximal inequality for martingale sequences [Lemma 3.7.1], we get

$$\mathbf{P}\left[\sup_{0\le t\le T}|V^r(t)-\widetilde{W}^r(t)|>\varepsilon\right]\le\frac{1}{(\varepsilon\zeta_r)^p}\,\mathbf{E}\left[\left|\sum_{k=1}^{[rT]}(Z_k^r-\mathbf{E}\,[Z_k^r])\right|^p\right] \tag{3.58}$$

with p selected such that $1 < p < \alpha \le 2$. This selection of p ensures $\mathbf{E}\,[\sigma^p] < \infty$ both under short–range dependence and under the assumption of regularly varying tail (3.9). The von Bahr – Esseen inequality [Lemma 3.7.2] for martingale differences can now be applied to the right–hand side of (3.58) to yield

$$\frac{1}{(\varepsilon\zeta_r)^p}\,\mathbf{E}\left[\left|\sum_{k=1}^{[rT]}(Z_k^r-\mathbf{E}\,[Z_k^r])\right|^p\right]\le\frac{2[rT]}{(\varepsilon\zeta_r)^p}\mathbf{E}\,[|Z^r-\mathbf{E}\,[Z^r]\,|^p]\,. \tag{3.59}$$

By the convexity of $x \to x^p$ $(p > 1)$ on $\mathbb{R}_+$, we find

$$\mathbf{E}\,[|Z^r-\mathbf{E}\,[Z^r]\,|^p]\le 2^{p-1}(\mathbf{E}\,[|Z^r|^p]+\mathbf{E}\,[|Z^r|]^p)\le 2^p\mathbf{E}\,[|Z^r|^p] \tag{3.60}$$

with the last step validated by Jensen's inequality. Next, using the definition of Z^r, we obtain by the same convexity argument that

$$\mathbf{E}\,[|Z^r|^p|\beta]\le\beta^{p-1}\,\mathbf{E}\left[\sum_{j=1}^{\beta}(1-U_j^r)^p\sigma_j^p\;|\beta\right]=\beta^p\mathbf{E}\,[\sigma^p]\,\mathbf{E}\,[(1-U^r)^p]\quad a.s. \tag{3.61}$$

under the enforced independence assumptions (and with an obvious notation). Injecting the bounds (3.60) and (3.61) into (3.59), we conclude from (3.58) that

$$\mathbf{P}\left[\sup_{0\le t\le T}|V^r(t)-\widetilde{W}^r(t)|>\varepsilon\right]\le\frac{2^{p+1}}{\varepsilon^p\zeta_r^{p-1}}\;\frac{[rT]}{r}\mathbf{E}\,[\sigma^p]\,\mathbf{E}\,[\beta^p]\,\frac{r}{\zeta_r}\mathbf{E}\,[(1-U^r)^p] \tag{3.62}$$

As the heavy traffic Assumption (B) implies

$$\lim_{r\to\infty}\frac{r}{\zeta_r}\mathbf{E}\,[(1-U^r)^p]=\frac{\gamma}{c}, \tag{3.63}$$

the desired conclusion (3.55) now follows by letting r go to infinity in (3.62) and noting that $\lim_{r\to\infty} 1/\zeta_r^{p-1} = 0$ for $p > 1$. ■

3.8 A proof of Proposition 3.6.1

In the proof of Proposition 3.6.1 and elsewhere, we make use of the following fact.

Lemma 3.8.1 *For any slowly varying function $u : \mathbb{R}_+ \to \mathbb{R}_+$, it holds that*

$$\lim_{x\to\infty} x^\rho u(x) = \infty, \quad \rho > 0, \tag{3.64}$$

while

$$\lim_{x\to\infty} x^\rho u(x) = 0, \quad \rho < 0. \tag{3.65}$$

Proof. By the Representation Theorem for slowly varying functions [7, Theorem 1.3.1, p. 12], we can write

$$u(x) \sim c \exp\left(\int_A^x \frac{\varepsilon(t)}{t}\, dt\right) \quad (x \to \infty) \tag{3.66}$$

with constants $A > 0$ and $c > 0$, and Borel mapping $\varepsilon : \mathbb{R}_+ \to \mathbb{R}$ such that $\lim_{t\to\infty} \varepsilon(t) = 0$. Thus,

$$x^\rho u(x) \sim cA^\rho \exp\left(\int_A^x \frac{\varepsilon(t)+\rho}{t}\, dt\right) \quad (x \to \infty).$$

For every $\delta > 0$ there exists $t_\delta > A$ such that $|\varepsilon(t)| < \delta$ for $t > t_\delta$, whence

$$\frac{-\delta+\rho}{t} \le \frac{\varepsilon(t)+\rho}{t} \le \frac{\delta+\rho}{t}, \quad t > t_\delta$$

so that

$$K + (-\delta+\rho)\ln\left(\frac{x}{t_\delta}\right) \le \int_A^x \frac{\varepsilon(t)+\rho}{t}\, dt \le K + (\delta+\rho)\ln\left(\frac{x}{t_\delta}\right), \quad x > t_\delta$$

with

$$K := \int_A^{t_\delta} \frac{\varepsilon(t)+\rho}{t}\, dt.$$

The conclusion (3.64) (resp. (3.65)) follows from these inequalities when selecting $\delta > 0$ such that $\delta < \rho$ (resp. $\delta < -\rho$) – such a selection is always possible when $\rho > 0$ (resp. $\rho < 0$). ■

The limit (3.64) is useful in the proof of the following discrete analogue to the direct half of Karamata's Theorem [7, p. 26].

Proposition 3.8.1 *Let $u : \mathbb{R}_+ \to \mathbb{R}_+$ be a slowly varying function. Then the following statements hold:*

(i) For any $p > -1$, we have the asymptotics

$$\sum_{n=1}^{r} n^p u(n) \sim \frac{r^{p+1}}{p+1}\, u(r) \quad (r \to \infty); \tag{3.67}$$

(ii) For any $p < -1$, we have the asymptotics

$$\sum_{n=r}^{\infty} n^p u(n) \sim -\frac{r^{p+1}}{p+1}\, u(r) \quad (r \to \infty); \tag{3.68}$$

(iii) The mapping $\widehat{u} : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$\widehat{u}(x) := \sum_{n=1}^{[x]} \frac{u(n)}{n}, \quad x \geq 0 \tag{3.69}$$

is a slowly varying function which satisfies

$$\lim_{x\to\infty} \frac{u(x)}{\widehat{u}(x)} = 0. \tag{3.70}$$

Proof. (i) Under the condition $p + 1 > 0$, it follows from (3.64) that (3.67) is equivalent to

$$\lim_{r\to\infty} \sum_{n=r^\star+1}^{r} \frac{1}{r} \left(\frac{n}{r}\right)^p \frac{u(n)}{u(r)} = \frac{1}{p+1} \tag{3.71}$$

for any finite $r^\star$ (which is now fixed for the remainder of the proof).

To prove (3.71) pick δ in $(0, 1+p)$ and $A > 0$. By Potter's bound [7, p. 25], there exists an integer $r^\star := r(\delta, A)$ such that

$$\frac{u(n)}{u(r)} < A\left(\frac{n}{r}\right)^{-\delta}, \quad r^\star \le n \le r. \tag{3.72}$$

Now, pick $r > r^\star$ and note that

$$\sum_{n=r^\star+1}^{r} \frac{1}{r}\left(\frac{n}{r}\right)^{p} \frac{u(n)}{u(r)} = \int_0^1 U_r(x)\, dx \tag{3.73}$$

where the function $U_r : \mathbb{R}_+ \to \mathbb{R}$ is defined by

$$U_r(x) := \mathbf{1}\left[x \ge \frac{r^\star}{r}\right] T_r(x)^p \, \frac{u(rT_r(x))}{u(r)}, \quad x \ge 0 \tag{3.74}$$

with

$$T_r(x) := \frac{[rx]+1}{r}, \quad x \ge 0. \tag{3.75}$$

For every x in $(0,1)$, we have $x \le T_r(x) \le 1$, and the Uniform Convergence Theorem for slowly varying functions [7, Theorem 1.2.1, p. 6] thus implies

$$\lim_{r\to\infty} \frac{u(rT_r(x))}{u(r)} = 1. \tag{3.76}$$

The pointwise convergence

$$\lim_{r\to\infty} U_r(x) = x^p \tag{3.77}$$

is now an immediate consequence of (3.74)–(3.76). Moreover, making use once more of Potter's bound (3.72) we find that

$$0 \le U_r(x) \le AT_r(x)^{p-\delta} \le A\max(1, x^{p-\delta})$$

with finite integral

$$\int_0^1 \max(1, x^{p-\delta})\, dx < 1 + \frac{1}{p-\delta+1} \tag{3.78}$$

under the choice of δ. These remarks together with (3.76) lead to

$$\lim_{r\to\infty}\int_0^1 U_r(x)dx = \int_0^1 x^p dx = \frac{1}{p+1} \tag{3.79}$$

by dominated convergence, and the desired limit (3.71) follows by going to the limit in (3.73).

(ii) The proof is similar to that for Part (i). We note that under the condition $p+1<0$ it follows from (3.65) that (3.68) is equivalent to

$$\lim_{r\to\infty}\sum_{n=r}^{\infty}\frac{1}{r}\left(\frac{n}{r}\right)^p\frac{u(n)}{u(r)} = -\frac{1}{p+1}. \tag{3.80}$$

Thus, pick δ in $(0,-1-p)$, and $A>0$. By a Potter bound [7, p. 25] similar to (3.72) there exists an integer $r^\star := r(\delta,A)$ such that

$$\frac{u(n)}{u(r)} < A\left(\frac{n}{r}\right)^\delta, \quad r^\star \le r \le n, \tag{3.81}$$

and we conclude that $\displaystyle\sum_{n=r}^{\infty} n^p u(n)$ is finite for each $r=1,2,\ldots$, under the current choice of δ. Next, we write

$$\sum_{n=r+1}^{\infty}\frac{1}{r}\left(\frac{n}{r}\right)^p\frac{u(n)}{u(r)} = \int_1^\infty U_r(x)\,dx \tag{3.82}$$

where the function $U_r : \mathbb{R}_+ \to \mathbb{R}$ is now defined by

$$U_r(x) := T_r(x)^p\ \frac{u(rT_r(x))}{u(r)}, \quad x\ge 0 \tag{3.83}$$

with $T_r(x)$ as in (3.75). Since, for every x in $[1,\infty)$, we have $x \le T_r(x) \le x+1$, the Uniform Convergence Theorem for slowly varying functions [7, Theorem 1.2.1, p. 6] implies (3.76) and the pointwise convergence (3.77) again follows. Making use of Potter's bound (3.81) we find that

$$0 \le U_r(x) \le AT_r(x)^{p+\delta} \le Ax^{p+\delta}, \quad r\ge r^\star$$

with finite integral

$$\int_1^\infty x^{p+\delta}\,dx = -\frac{1}{p+\delta+1} \tag{3.84}$$

under the current choice of δ. These remarks together with (3.77) lead to

$$\lim_{r\to\infty}\int_1^\infty U_r(x)dx = \int_1^\infty x^p dx = -\frac{1}{p+1} \tag{3.85}$$

by dominated convergence, and the desired limit (3.80) follows by going to the limit in (3.82).

(iii) We begin by noting that the standard asymptotics

$$\lim_{x\to\infty}\sum_{n=[ax]+1}^{[bx]}\frac{1}{n} = \ln\left(\frac{b}{a}\right), \quad 0<a<b$$

imply

$$\lim_{x\to\infty}\sum_{n=[ax]+1}^{[bx]}\frac{1}{n}\frac{u(n)}{u(x)} = \ln\left(\frac{b}{a}\right), \quad 0<a<b. \tag{3.86}$$

Indeed, for some $x^{\star\star}$ in $\mathbb{R}_+$, it holds that

$$a \le \sup\left(\frac{n}{x}:\ n=[ax]+1,\dots,[bx]\right) \le b, \quad x\ge x^{\star\star}. \tag{3.87}$$

By the Uniform Convergence Theorem [7, Theorem 1.2.1, p. 6], for each $\delta>0$ there exists $x^\star := x_{a,b}(\delta) > x^{\star\star}$ such that

$$\sup\left(\left|\frac{u(tx)}{u(x)}-1\right|:\ t\in[a,b]\right) \le \delta, \quad x\ge x^\star. \tag{3.88}$$

Combining (3.87) and (3.88) we readily get (3.86) in the form

$$\lim_{x\to\infty}\sum_{n=[ax]+1}^{[bx]}\frac{1}{n}\frac{u(\frac{n}{x}x)}{u(x)} = \ln\left(\frac{b}{a}\right).$$

Now, pick ε in $(0,1)$. In view of (3.86) we have

$$\liminf_{x\to\infty}\frac{\widehat{u}(x)}{u(x)} \ge \lim_{x\to\infty}\sum_{n=[\varepsilon x]+1}^{[x]}\frac{1}{n}\frac{u(n)}{u(x)} = -\ln\varepsilon. \tag{3.89}$$

Since ε can be chosen arbitrarily small, it follows that

$$\lim_{x\to\infty}\frac{\widehat{u}(x)}{u(x)}=\infty \tag{3.90}$$

or equivalently, (3.70). To prove that $\widehat{u}$ is slowly varying, pick $y>1$ and note that for every $x>0$, we have

$$\widehat{u}(yx)=\widehat{u}(x)+u(x)\sum_{n=[x]+1}^{[yx]}\frac{1}{n}\frac{u(n)}{u(x)},$$

so that

$$\frac{\widehat{u}(yx)}{\widehat{u}(x)}=1+\frac{u(x)}{\widehat{u}(x)}\sum_{n=[x]+1}^{[yx]}\frac{1}{n}\frac{u(n)}{u(x)}.$$

It is now straightforward from (3.70) and (3.86) to obtain

$$\lim_{x\to\infty}\frac{\widehat{u}(yx)}{\widehat{u}(x)}=1.$$

The case $y<1$ is handled in a similar way, and the slow variation of $\widehat{u}$ follows. ■

A proof of Proposition 3.6.1.

From (2.3) it always holds that

$$\mathbf{E}\left[\widehat{\sigma}\right]=\sum_{n=0}^{\infty}\mathbf{P}\left[\widehat{\sigma}>n\right]=\frac{\mathbf{E}\left[\sigma^2\right]}{2\mathbf{E}\left[\sigma\right]}+\frac{1}{2}. \tag{3.91}$$

We consider each of the scalings $\{\zeta_r,\ r=1,2,\ldots\}$ associated with Theorems 3.3.1 – 3.3.3, separately:

[Theorem 3.3.1] Under short–range dependence, we have $\mathbf{E}\left[\sigma^2\right]<\infty$, and it is immediate from (3.91) that Condition (B) holds for the choice $\zeta_r=\sqrt{r}$, $r=1,2,\ldots$ (in fact for any choice such that $\lim_{r\to\infty}\zeta_r=\infty$). ■

We next turn to Theorems 3.3.2 and 3.3.3. Upon substituting (3.9) (with $1<\alpha\leq$

2) into (2.3), we readily get from Proposition 3.8.1(ii) that

$$\mathbf{P}\left[\widehat{\sigma} > n\right] = \frac{1}{\mathbf{E}\left[\sigma\right]} \sum_{j=n}^{\infty} \mathbf{P}\left[\sigma > j\right] \sim \frac{1}{(\alpha-1)\mathbf{E}\left[\sigma\right]}\, n^{1-\alpha} h(n) \quad (n \to \infty),$$

whence

$$\sum_{n=1}^{r} \mathbf{P}\left[\widehat{\sigma} > n\right] \sim \frac{1}{(\alpha-1)\mathbf{E}\left[\sigma\right]} \sum_{n=1}^{r} n^{1-\alpha} h(n) \quad (r \to \infty) \tag{3.92}$$

provided $\mathbf{E}\left[\widehat{\sigma}\right]$ is infinite.

[Theorem 3.3.2] When $\alpha = 2$ in (3.9), the condition $\mathbf{E}\left[\sigma^2\right] = \infty$ implies that $\mathbf{E}\left[\widehat{\sigma}\right]$ is infinite by (3.91). Thus, (3.92) holds in the form

$$\sum_{n=1}^{r} \mathbf{P}\left[\widehat{\sigma} > n\right] \sim \frac{1}{\mathbf{E}\left[\sigma\right]} \sum_{n=1}^{r} \frac{h(n)}{n} \quad (r \to \infty)$$

which, from Proposition 3.8.1(iii) is seen to be slowly varying. By Proposition 3.3.1, the scaling $\{\zeta_r,\ r = 1, 2, \ldots\}$ is $1/2$–regularly varying, so that

$$\frac{1}{\zeta_r} \sum_{n=1}^{r} \mathbf{P}\left[\widehat{\sigma} > n\right] \sim \frac{1}{\mathbf{E}\left[\sigma\right]} r^{-\frac{1}{2}} \frac{1}{\widehat{h}(r)} \sum_{n=1}^{r} \frac{h(n)}{n} \quad (r \to \infty) \tag{3.93}$$

for some slowly varying function $\widehat{h} : \mathbb{R}_+ \to \mathbb{R}_+$. The ratio of slowly varying functions being itself slowly varying, we readily conclude from Lemma 3.8.1 and (3.93) that Condition (B) holds. ■

[Theorem 3.3.3] On the range $1 < \alpha < 2$, $\mathbf{E}\left[\sigma^2\right]$ is infinite, and so is $\mathbf{E}\left[\widehat{\sigma}\right]$ by virtue of (3.91). Proposition 3.8.1(i) applied to the right–hand side of the asymptotic equivalence (3.92) yields

$$\sum_{n=1}^{r} \mathbf{P}\left[\widehat{\sigma} > n\right] \sim \frac{1}{(2-\alpha)(\alpha-1)\mathbf{E}\left[\sigma\right]}\, r^{2-\alpha} h(r) \quad (r \to \infty). \tag{3.94}$$

By Proposition 3.3.1 the scaling $\{\zeta_r,\ r = 1, 2, \ldots\}$ is $1/\alpha$–regularly varying, so that

$$\frac{1}{\zeta_r} \sum_{n=1}^{r} \mathbf{P}\left[\widehat{\sigma} > n\right] \sim \frac{1}{(2-\alpha)(\alpha-1)\mathbf{E}\left[\sigma\right]}\, \frac{h(r)}{\widehat{h}(r)} \frac{r^{2-\alpha}}{r^{\frac{1}{\alpha}}} \quad (r \to \infty) \tag{3.95}$$

for some slowly varying function $\widehat{h} : \mathbb{R}_+ \to \mathbb{R}_+$. The ratio of slowly varying functions is itself slowly varying, and Condition (B) is now a direct consequence of Lemma 3.8.1 once we note that $2 - \alpha - \alpha^{-1} < 0$ for $\alpha > 0$. ■

Chapter 4

Light traffic limits

4.1 Introduction

We seek to characterize the light traffic limiting behavior of the queueing system with $M|G|\infty$ inputs. Our main tool for accomplishing this task is a methodology presented in a series of papers by Reiman and Simon [49, 50, 51]. Their approach provides a general framework for deriving asymptotic results in systems where the quantity of interest can be expressed as a function of a Poisson–like process. If $\lambda > 0$ denotes the intensity of the Poisson process driving the system, then the light traffic information furnished by the Reiman–Simon technique consists of derivatives of the quantity of interest, with respect to λ, evaluated at $\lambda = 0+$. We devote most of our efforts to the case $c = 1$, because then additional expressions become available by relating the system with $M|G|\infty$ inputs to that with instantaneous inputs. The results quantify the effect of the session duration distribution G and reveal the differences between the gradual $M|G|\infty$ inputs and the instantaneous arrivals of a classical $GI|GI|1$ queue. Moreover, in the special case $c = 1$, they suggest a classification of the light traffic behavior of the multiplexer buffer content in terms of the short– vs long– range dependent property of the $M|G|\infty$ arrival

process.

4.2 Reiman–Simon theory

4.2.1 Preliminaries

To place our problem in the context of the Reiman–Simon methodology, we start by introducing bi–infinite counterparts to the sequences of $\mathbb{N}$–valued rvs representing session arrivals and their activity durations. That is, we consider the bi–infinite sequences of $\mathbb{N}$–valued rvs $\{\beta_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ and $\{\sigma_{n,j},\ n = 0, \pm 1, \pm 2, \ldots;\ j = 1, 2, \ldots\}$ which are constructed on a common probability space $(\Omega, \mathcal{F}, \mathbf{P}_\lambda)$, are mutually independent and satisfy Assumptions (A)(i) and (ii), respectively. We also introduce the bi–infinite sequences of $M|G|\infty$ inputs $\{b_n,\ n = 0, \pm 1, \pm 2, \ldots\}$ and queue lengths $\{q_n,\ n = 0, \pm 1, \pm 2, \ldots\}$, where the former are given by

$$b_n := \sum_{k=-\infty}^{n} \sum_{j=1}^{\beta_k} \mathbf{1}\left[\sigma_{k,j} > n - k\right], \quad n = 0, \pm 1, \ldots,$$

so that b_n denotes the number of active sessions at the beginning of slot $[n, n+1)$. In this setup, instead of starting the Lindley recursion at time $n = 0$ we take the viewpoint that the system has been operating from time $n = -\infty$, i.e.,

$$q_{n+1} = [q_n + b_{n+1} - c]^+, \quad n = 0, \pm 1, \ldots. \tag{4.1}$$

Under the stability condition $\lambda \mathbf{E}[\sigma] < c$ convergence to the stationary $\mathbb{R}_+$–valued rv q_∞ has already taken place by time $n = 0$, that is,

$$q_0 =_{st} q_\infty. \tag{4.2}$$

Application of the Reiman–Simon method entails conditioning on the number of arriving sessions and their corresponding activity durations. We introduce the

necessary notation. First, we denote by the empty set $\emptyset$ the event that, from time $t = -\infty$ to $t = +\infty$, there are no session arrivals at all. Next, fix $n = 1, 2, \ldots$, and consider the following event ω: There are exactly n sessions that ever become active. For each $i = 1, 2, \ldots, n$, fixed $t_i = 0, \pm 1, \ldots$, and $k_i = 1, 2, \ldots$, the i^{th} session becomes active during time slot time $[t_i - 1, t_i)$ and its activity period lasts k_i time slots. We denote such an event ω by $\{t_1, \ldots, t_n; k_1, \ldots, k_n\}$. In other words, $\{t_1, \ldots, t_n; k_1, \ldots, k_n\}$ corresponds to an event where n sessions arrive to the system over all time, and these arrivals occur during time slots $[t_1 - 1, t_1), \ldots, [t_n - 1, t_n)$ with respective activity durations $k_1, \ldots, k_n$. We drop the duration indices to denote unions of events, i.e.,

$$\{t_1, \ldots, t_n; \} := \bigcup_{k_1=1}^{\infty} \cdots \bigcup_{k_n=1}^{\infty} \{t_1, \ldots, t_n; k_1, \ldots, k_n\}, \tag{4.3}$$

is the event that exactly n sessions arrive to the system over all time, and these arrivals occur during time slots $[t_1 - 1, t_1), \ldots, [t_n - 1, t_n)$.

4.2.2 Light traffic derivatives

Let the generic system performance metric $\phi(\lambda)$ be expressed as

$$\phi(\lambda) = \int \psi \, d\mathbf{P}_\lambda \tag{4.4}$$

for a suitably chosen rv $\psi : \Omega \to \mathbb{R}$. For example, $\psi(\omega)$ can be chosen as the queue length at time $t = 0$, corresponding to a sample path ω in Ω, in which case, from (4.2) and (4.4), the performance metric $\phi(\lambda)$ is the expected value $\mathbf{E}_\lambda [q_\infty]$ of the stationary queue length q_∞.

Following the Reiman–Simon method, we decompose the expectation in (4.4) according to occurrences of session arrivals/activity durations events of the form $\{t_1, \ldots, t_n; k_1, \ldots, k_n\}$. To do this, for each $n = 1, 2, \ldots$ we associate with ψ several

auxiliary functions. First, the expected value $\widehat{\psi}$ of ψ, conditional on the session arrivals event $\{t_1, \ldots, t_n;\}$ is given by

$$\widehat{\psi}(\{t_1, \ldots, t_n;\}) := \mathbf{E}\left[\psi \mid \{t_1, \ldots, t_n;\}\right]; \tag{4.5}$$

this does not depend on λ. Next, we define the function $\widetilde{\psi} : \{1, 2, \ldots\}^n \to \mathbb{R}$ by

$$\widetilde{\psi}(k_1, \ldots, k_n) := \sum_{t_1=-\infty}^{+\infty} \cdots \sum_{t_n=-\infty}^{+\infty} \psi(\{t_1, \ldots t_n; k_1, \ldots, k_n\}). \tag{4.6}$$

where $\psi(\{t_1, \ldots t_n; k_1, \ldots, k_n\})$ is the value of ψ when n sessions arrive to the system over all time, these session arrivals occur during time slots $[t_1 - 1, t_1), \ldots, [t_n - 1, t_n)$, and their respective activity durations are given by $k_1, \ldots, k_n$. Furthermore, let $\Pi_{n,j}$ denote the set of unordered j-tuples chosen from $\{1, 2, \ldots, n\}$ (with repetitions allowed), where for $\pi = \{i_1, i_2, \ldots, i_j\}$ in $\Pi_{n,j}$, we use the notation $\mathbf{t}_\pi := \{t_{i_1}, t_{i_2}, \ldots, t_{i_j};\}$.

Now, for any given arrival event $\{t_1, \ldots, t_n;\}$, we define

$$\Psi(\{t_1, \ldots, t_n;\}) := \sum_{j=0}^{n} (-1)^{n-j} \sum_{\pi \in \Pi_{n,j}} \widehat{\psi}(\mathbf{t}_\pi). \tag{4.7}$$

For instance, we have

$$\Psi(\{t;\}) = \widehat{\psi}(\{t;\}) - \widehat{\psi}(\emptyset), \tag{4.8}$$

$$\Psi(\{t_1, t_2;\}) = \widehat{\psi}(\{t_1, t_2;\}) - \widehat{\psi}(\{t_1;\}) - \widehat{\psi}(\{t_2;\}) + \widehat{\psi}(\emptyset) \tag{4.9}$$

and so on.

The formulas for the light traffic derivatives can be obtained by considering a system where only arrivals in an interval of the form $[-T, T)$, for $T = 1, 2, \ldots$, are ever allowed to enter; let $\phi^T(\lambda)$ be the corresponding performance metric. The idea is to calculate first the derivatives of $\phi^T(\lambda)$ with respect to λ, at $\lambda = 0+$, and then let T go to infinity. Clearly, it is necessary to justify that this interchange of limits

in λ and T leads to the correct answer. To that end we enforce an assumption on the finiteness of the exponential moment of σ:

Assumption (C) *There exists $\theta^\star > 0$ such that $\mathbf{E}\left[e^{\theta\sigma}\right] < \infty$ for $\theta < \theta^\star$.*

In [51] it is shown that under Assumption (C) the interchange of limits is indeed valid; here we simply restate this conclusion as

Proposition 4.2.1 *Under Assumption (C) it holds that*

$$\lim_{T\to\infty} \frac{d^n}{d\lambda^n}\ \phi^T(0+) = \frac{d^n}{d\lambda^n}\ \phi(0+), \quad n = 0, 1, \ldots.$$

The following result is essentially a discrete–time version of Theorem 2 in [51, p. 30], and enables us to calculate the n^{th} order derivative of $\phi(\lambda)$ at $\lambda = 0+$ by considering scenarios where at most n sessions ever arrive to the system.

Proposition 4.2.2 *If Assumption (C) is satisfied, then*

$$\lim_{\lambda\to 0+} \phi(\lambda) = \widehat{\psi}(\emptyset), \tag{4.10}$$

and for each $n = 1, 2, \ldots$, it holds that

$$\frac{d^n}{d\lambda^n}\ \phi(0+) = \sum_{t_1=-\infty}^{+\infty} \cdots \sum_{t_n=-\infty}^{+\infty} \Psi(\{t_1, \ldots, t_n;\}). \tag{4.11}$$

Proof. For each $T = 1, 2, \ldots$ and $j = 0, 1, \ldots$ let

$$P_j^T(\lambda) := e^{-2\lambda T} \frac{(2\lambda T)^j}{j!}$$

denote the probability that j discrete–time Poisson session arrivals occur during the interval $[-T, T)$. For each $n = 0, 1, \ldots$, the n^{th} derivative of $P_j^T(\lambda)$ with respect to λ is given by

$$\frac{d^n}{d\lambda^n}\ P_j^T(\lambda) = (2T)^n \sum_{i=0}^{\min(n,j)} (-1)^{n-i} \left(\begin{array}{c} n \\ i \end{array} \right) P_{j-i}^T(\lambda),$$

so that

$$\frac{d^n}{d\lambda^n}\ P_j^T(0+) = \begin{cases} (2T)^n \begin{pmatrix} n \\ j \end{pmatrix} (-1)^{n-j} & \text{if } \ n \geq j \\ 0 & \text{if } \ n < j. \end{cases} \tag{4.12}$$

Given that j Poisson arrivals have occured in $[-T, T)$, they are uniformly distributed over the $2T$ time slots. Thus,

$$\phi^T(\lambda) = \sum_{j=0}^{\infty} P_j^T(\lambda) \sum_{t_1=-T+1}^{T} \cdots \sum_{t_j=-T+1}^{T} \frac{1}{(2T)^j}\ \widehat{\psi}(\{t_1, \ldots, t_j;\})$$

and using (4.12) we get

$$\begin{aligned} \frac{d^n}{d\lambda^n}\ \phi^T(0+) &= \sum_{j=0}^{n} (2T)^n \begin{pmatrix} n \\ j \end{pmatrix} (-1)^{n-j} \sum_{t_1=-T+1}^{T} \cdots \sum_{t_j=-T+1}^{T} \frac{1}{(2T)^j}\ \widehat{\psi}(\{t_1, \ldots, t_j;\}) \\ &= \sum_{j=0}^{n} (2T)^{n-j} (-1)^{n-j} \sum_{\pi \in \Pi_{n,j}} \sum_{\mathbf{t}_\pi=-T+1}^{T} \widehat{\psi}(\mathbf{t}_\pi) \\ &= \sum_{j=0}^{n} (-1)^{n-j} \sum_{\pi \in \Pi_{n,j}} \sum_{t_1=-T+1}^{T} \cdots \sum_{t_n=-T+1}^{T} \widehat{\psi}(\mathbf{t}_\pi) \\ &= \sum_{t_1=-T+1}^{T} \cdots \sum_{t_n=-T+1}^{T} \Psi(\{t_1, \ldots, t_n;\}) \end{aligned} \tag{4.13}$$

Letting T go to infinity in (4.13) and invoking Proposition 4.2.1 we conclude that (4.11) holds true. ■

We rely on Proposition 4.2.2 to calculate light traffic derivatives of system quantities in the sequel.

4.2.3 Case $c \geq 1$

We now consider a Lindley recursion (4.1) with release rate $c \geq 1$. Fix some integer $p = 1, 2, \ldots$. Take $\psi := {q_0}^p$, where q_0 is the queue length at time $n = 0$, so that the performance measure of interest is $\phi(\lambda) = \mathbf{E}_\lambda\left[{q_\infty}^p\right]$, the p^{th} moment of the

stationary queue size. To determine its light traffic derivatives we need to evaluate the quantities appearing in Proposition 4.2.2.

Clearly, if at most $\lfloor c \rfloor$ sessions are ever active, then, since each active session generates one arrival per time slot, their inputs are flushed out of the queue by the end of the time slot and the queue remains empty. So, for each $m = 1, 2, \ldots, \lfloor c \rfloor$ and $t_1, \ldots, t_m = 0, \pm 1, \pm 2, \ldots$ it is immediate from definition (4.5) that

$$\widehat{\psi}(\emptyset) = 0 \quad \text{and} \quad \widehat{\psi}(\{t_1, \ldots, t_m; \}) = 0. \tag{4.14}$$

Next, fix some $b \geq 0$ and take $\psi := \mathbf{1}[q_0 > b]$, in which case the performance metric of interest is the tail probability, $\phi(\lambda) = \mathbf{E}_\lambda[\psi] = \mathbf{P}_\lambda[q_\infty > b]$. If at most $\lfloor c \rfloor$ sessions become active, then the same simple considerations as before apply, because whenever $q_0 = 0$ we also have $\mathbf{1}[q_0 > b] = 0$. Thus relations (4.14) still hold true for the current choice $\psi := \mathbf{1}[q_0 > b]$ as well. Consequently, by combining (4.14) with Proposition 4.2.2 for each of the functions $\psi := {q_0}^p$ and $\psi := \mathbf{1}[q_0 > b]$ we arrive at

Proposition 4.2.3 *Consider a Lindley recursion (4.1) with release rate $c \geq 1$. If Assumption (C) is satisfied then for each $m = 1, 2, \ldots, \lfloor c \rfloor$, it holds that:*

(a) *For $p = 1, 2, \ldots$,*

$$\lim_{\lambda \to 0+} \mathbf{E}_\lambda[{q_\infty}^p] = 0 \quad \textit{and} \quad \frac{d^m}{d\lambda^m}\, \mathbf{E}_\lambda[{q_\infty}^p]\Big|_{\lambda=0+} = 0. \tag{4.15}$$

(b) *For $b \geq 0$,*

$$\lim_{\lambda \to 0^+} \mathbf{P}_\lambda[q_\infty > b] = 0 \quad \textit{and} \quad \frac{d^m}{d\lambda^m}\, \mathbf{P}_\lambda[q_\infty > b]\Big|_{\lambda=0+} = 0. \tag{4.16}$$

Before proceeding we make a few comments. Note that when $c = 1$ the server can only process one session at a time, so that the system can be viewed as a single server queue operating in discrete time. Then from (4.16) with $c = 1$ it is already

apparent that the light traffic limits of a queueing system with $M|G|\infty$ inputs differ from those of a standard single server $GI|GI|1$ queue. For the system first derivative is here zero, while in the system (2.17) with instantaneous inputs the first light traffic derivative is positive. This is a manifestation of the fact that work that joins the system gradually, as is the case with $M|G|\infty$ inputs, generates less queueing that work arriving instantaneously. Also, relation (4.16) reflects (though in a rough manner) the statistical multiplexing gain: All powers of λ up to and including $\lambda^{\lfloor c \rfloor}$ offer no contribution to the tail probability. Thus (4.16) implies that, in light traffic, increasing the multiplexer release rate c while maintaining the same system utilization $\lambda \mathbf{E}[\sigma]/c$ results in a decreasing tail probability $\mathbf{P}_\lambda[q_\infty > b]$.

In view of (4.15) and (4.16) the focus shifts to the calculation of the derivative of order $\lfloor c \rfloor + 1$. This is the first non-zero derivative and it is clearly more informative than the $\lfloor c \rfloor + 1$ lower order derivatives, for it provides the leading term in expansions of system quantities around $\lambda = 0$.

4.2.4 Case $c = 1$

In this section we consider a Lindley recursion (4.1) with release rate $c = 1$. This corresponds to the situation where each active session in the $M|G|\infty$ input generates arrivals at rate equal to the multiplexer service rate. In this case we are able to obtain explicit expressions for the second order light traffic derivatives of system quantities by carrying out in full the calculations associated with Proposition 4.2.2.

We take ψ to be either $\psi := {q_0}^p$ for some $p = 1, 2, \ldots$, or $\psi := \mathbf{1}[q_0 > b]$ for $b \geq 0$, yielding $\phi(\lambda) = \mathbf{E}_\lambda[{q_\infty}^p]$ and $\phi(\lambda) = \mathbf{P}_\lambda[q_\infty > b]$, respectively. From (4.11)

it follows that

$$\frac{d^2}{d\lambda^2}\,\phi(0+) = \sum_{t_1=-\infty}^{+\infty}\sum_{t_2=-\infty}^{+\infty}\widehat{\psi}(\{t_1,t_2;\}), \tag{4.17}$$

where we have also taken into account (4.14). We thus need to evaluate $\widehat{\psi}(\{t_1,t_2;\})$. This calculation requires consideration of a system where only two sessions are ever active. In particular, since

$$\widehat{\psi}(\{t_1,t_2;\}) = \sum_{k_1=1}^{\infty}\sum_{k_2=1}^{\infty}\psi(\{t_1,t_2;k_1,k_2\})\mathbf{P}\left[\sigma=k_1\right]\mathbf{P}\left[\sigma=k_2\right], \tag{4.18}$$

we need only examine the queue length process induced by events of the form $\{t_1,t_2;k_1,k_2\}$. It is convenient to interchange the order of summations appearing in (4.17) and (4.18) – this can be done without qualms for the summands are all non-negative. Recalling definition (4.6), we calculate the sums over the arrival times first, say

$$\widetilde{\psi}(k_1,k_2) = \sum_{t_1=-\infty}^{+\infty}\sum_{t_2=-\infty}^{+\infty}\psi(\{t_1,t_2;k_1,k_2\}), \tag{4.19}$$

and then rewrite (4.17) as

$$\frac{d^2}{d\lambda^2}\,\phi(0+) = \mathbf{E}\left[\widetilde{\psi}(\sigma_1,\sigma_2)\right], \tag{4.20}$$

where σ_1 and σ_2 are i.i.d. copies of the generic activity duration rv σ. Thus, the calculation of the second derivative is reduced to the evaluation of $\widetilde{\psi}$. To determine $\widetilde{\psi}$ consider a session arrival/activity duration event of the form $\{t_1,t_2;k_1,k_2\}$. Observe that if

$$\min(t_2+k_2,t_1+k_1) > \max(t_1,t_2) \tag{4.21}$$

the two sessions that arrive in slots $[t_1-1,t_1)$ and $[t_2-1,t_2)$ are simultaneously active from time $\max(t_1,t_2)$ until $\min(t_2+k_2,t_1+k_1)$. In that case the queue size evolves as shown in Figure 4.1; otherwise, i.e., if $\min(t_2+k_2,t_1+k_1) \leq \max(t_1,t_2)$, it is identically zero.

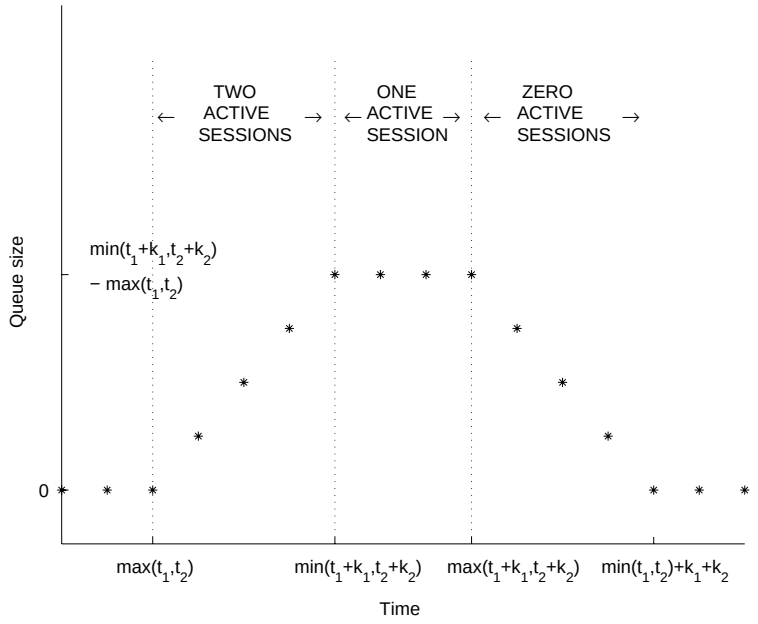

Figure 4.1: Queue length evolution under the event $\{t_1, t_2; k_1, k_2\}$

By inspection, under (4.21), the queue length at time $t = 0$ is given by

$$q_0(\{t_1,t_2;k_1,k_2\}) = \begin{cases} -\max(t_1,t_2), & \text{if } \max(t_1,t_2) \le 0 \text{ and } 0 \le \min(t_1+k_1,t_2+k_2) \\ \min(t_1+k_1,t_2+k_2) & \text{if } \min(t_1+k_1,t_2+k_2) < 0 \\ \quad -\max(t_1,t_2), & \text{and } 0 \le \max(t_1+k_1,t_2+k_2) \\ \min(t_1,t_2)+k_1+k_2, & \text{if } \max(t_1+k_1,t_2+k_2) < 0 \text{ and } 0 \le \min(t_1,t_2)+k_1+k_2 \\ 0, & \text{otherwise.} \end{cases}$$

To facilitate the evaluation of $\widehat{\psi}$ we display $q_0(\{t_1,t_2;k_1,k_2\})$ in the t_1t_2–plane; the values for $k_1 \ge k_2$ and $k_1 < k_2$ are shown in Figures 4.2 and 4.3, respectively.

We carry out the detailed calculation of the function $\widetilde{\psi}$ corresponding to each of the choices $\psi := q_0$, $\psi := {q_0}^2$ and $\psi := \mathbf{1}[q_0 > b]$, $b = 0, 1, \ldots$, in Section 4.4. The results are summarized in Lemma 4.2.1 below.

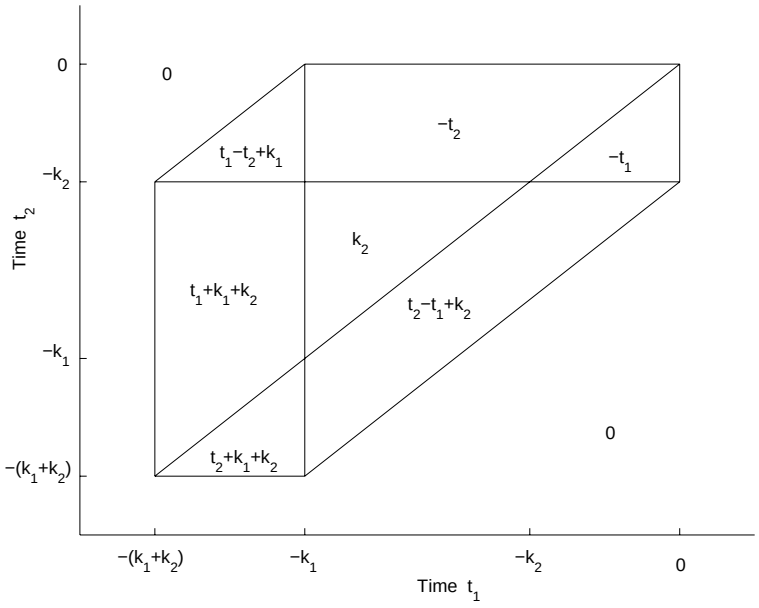

Figure 4.2: Values of $q_0(\{t_1, t_2; k_1, k_2\})$ when $k_1 \geq k_2$.

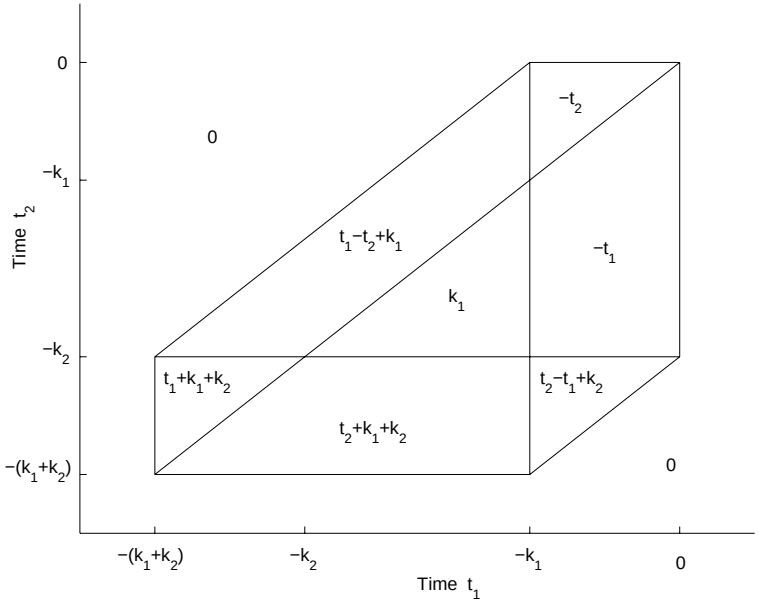

Figure 4.3: Values of $q_0(\{t_1, t_2; k_1, k_2\})$ when $k_1 < k_2$.

Lemma 4.2.1 *Assume c=1 in the Lindley recursion (4.1). The following statements hold:*

(a) *If* $\psi := q_0$ *then*

$$\widetilde{\psi}(k_1, k_2) = \frac{1}{2}\, k_1 k_2 (k_1 + k_2), \quad k_1, k_2 = 1, 2, \ldots. \tag{4.22}$$

(b) *If* $\psi := {q_0}^2$ *then*

$$\widetilde{\psi}(k_1, k_2) = \frac{1}{2}\, k_1 k_2 (1 + k_1 k_2), \quad k_1, k_2 = 1, 2, \ldots. \tag{4.23}$$

(c) *If* $\psi := \mathbf{1}\left[q_0 > b\right]$ *for some* $b = 0, 1, \ldots,$ *then*

$$\begin{aligned} \widetilde{\psi}(k_1, k_2) \;\; &= \;\; \frac{1}{2}\left\{(k_1 - b)^2 + (k_2 - b)^2\right\} - \frac{3}{2}\,(k_1 + k_2 - 2b) \\ &\quad +2(k_1 - b)(k_2 - b) + 1 \end{aligned} \tag{4.24}$$

for $k_1, k_2 > b$ *and* $\widetilde{\psi}(k_1, k_2) = 0$ *otherwise.*

In principle it is possible to evaluate the function $\widetilde{\psi}$ corresponding to $\psi := {q_0}^p$ for $p > 2$ in a similar manner. However, such calculations become increasingly tedious for large p.

We now obtain explicit expressions for the second order light traffic derivatives of system quantities by combining each one of (4.22), (4.23) and (4.24) with (4.20). This leads to the following

Proposition 4.2.4 *Let* $c = 1$ *in the Lindley recursion (4.1). Under Assumption (C) it holds that:*

(a) *The moments* $\mathbf{E}_\lambda\left[q_\infty\right]$ *and* $\mathbf{E}_\lambda\left[{q_\infty}^2\right]$ *satisfy*

$$\frac{d^2}{d\lambda^2}\; \mathbf{E}_\lambda\left[q_\infty\right]\Big|_{\lambda=0+} = \mathbf{E}\left[\sigma\right]\mathbf{E}[\sigma^2] \tag{4.25}$$

and

$$\frac{d^2}{d\lambda^2}\; \mathbf{E}_\lambda\left[{q_\infty}^2\right]\Big|_{\lambda=0+} = \frac{1}{2}\left(\mathbf{E}\left[\sigma\right]^2 + \mathbf{E}[\sigma^2]^2\right). \tag{4.26}$$

(b) *For each* $b = 0, 1, \ldots$

$$\frac{d^2}{d\lambda^2}\,\mathbf{P}_\lambda\left[q_\infty > b\right]\Big|_{\lambda=0+} = \mathbf{E}\left[(\sigma-b)^{+2}\right]\mathbf{P}\left[\sigma > b\right] + 2\,\mathbf{E}\left[(\sigma-b)^{+}\right]^2 - 3\,\mathbf{E}\left[(\sigma-b)^{+}\right]\mathbf{P}\left[\sigma > b\right] + \mathbf{P}\left[\sigma > b\right]^2. \quad (4.27)$$

Proposition 4.2.4 delineates a light traffic behavior for the queue with $M|G|\infty$ arrivals that is certainly different from that of a classical $GI|GI|1$ queue. Indeed, when considering the first two terms in a light traffic expansion of $\mathbf{P}_\lambda\left[q_\infty > b\right]$ around $\lambda = 0$ (for $c = 1$) we see that the first derivative (4.16) is zero so that the second derivative is the most informative. This is given by (4.27) which highlights the role of the activity duration rv σ, through both its distribution and its first two moments. Thus, a light traffic expansion of the tail probability $\mathbf{P}_\lambda\left[q_\infty > b\right]$ induced by $M|G|\infty$ arrivals is completely different from the corresponding expansion for the system with instantaneous inputs: This can be obtained from (2.38) (or via the Reiman–Simon method) and is given by $\mathbf{P}_\lambda\left[q_\infty^{(u)} > b\right] \sim \lambda\mathbf{E}\left[\sigma\right]\mathbf{P}\left[\widehat{\sigma} > b+1\right]$ $(\lambda \to 0)$, thus starting with a non-zero first order term λ. Notice also that here, even if Assumption (C) were to be relaxed, (4.27) shows that for $\mathbf{P}_\lambda\left[q_\infty > b\right]$ to decay like λ^2 for small λ it is necessary that $\mathbf{E}\left[\sigma^2\right]$ be finite. If $\mathbf{E}\left[\sigma^2\right] = \infty$, as is the case for long–range dependent $M|G|\infty$ arrivals, expression (4.27) yields infinity and λ^2 is no longer the correct order of decay. A different, smaller exponent should be sought in the long–range dependent case.

4.2.5 A heavy–light traffic relationship

In Sections 4.2.3 and 4.2.4 we have, at least partially, mapped out the light traffic behavior of the multiplexer with $M|G|\infty$ inputs. Of course this partial information is augmented whenever expressions for the next higher order derivatives become

available. Unfortunately, the calculations soon become intractable and, typically, explicit expressions are available only for the first non-zero derivative. In principle however, if light traffic derivatives of every order were known, then system quantities would be completely determined away from the light traffic regime by their Taylor series expansion (under analyticity assumptions). In particular, knowledge of all light traffic derivatives would also imply full knowledge of system quantities in heavy traffic. This observation raises the question as to how these successive derivatives at $\lambda = 0+$ are related to the respective limiting behavior of system quantities in the heavy traffic regime. The answer is given in [56], where a simple yet rather unexpected relationship is established. This relationship suggests a method for constructing certain approximations of system quantities; these are precisely the interpolation approximations discussed in Chapter 5.

The link between heavy traffic limits and light traffic derivatives is provided by the following proposition, which is a special case of Simon's results [56].

Proposition 4.2.5 *Let the function $\phi : [0,1) \to \mathbb{R}_+$ be analytic on $[0,1)$. Suppose that the limit*

$$H := \lim_{x \to 1-} (1-x)\phi(x) \tag{4.28}$$

exists and is positive and finite. Then

$$\lim_{n \to \infty} \frac{1}{n!} \frac{d^n}{dx^n} \phi(0+) = H. \tag{4.29}$$

We apply Proposition 4.2.5 in the context of the Lindley recursion (4.1) with $c = 1$, describing a system with $M|G|\infty$ arrivals, by setting $x := \lambda \mathbf{E}[\sigma]$, so that x lies in $[0,1)$ whenever the system is stable. Let us assume $\phi(\lambda) := \mathbf{E}_\lambda [q_\infty]$ is analytic in $[0,1)$. From Assumption (B) and Theorems 3.3.1 and 3.4.1 we infer that the limit

$$H := \lim_{\lambda \to 1/\mathbf{E}[\sigma]} (1 - \lambda \mathbf{E}[\sigma])\, \mathbf{E}_\lambda [q_\infty] = \frac{\mathbf{E}[\sigma^2]}{2\mathbf{E}[\sigma]} \tag{4.30}$$

is positive and finite whenever $\mathbf{E}\left[\sigma^2\right] < \infty$. In that case Proposition 4.2.5 reads

$$\lim_{n\to\infty} \frac{1}{n!} \frac{1}{\mathbf{E}\left[\sigma\right]^n} \frac{d^n}{d\lambda^n} \mathbf{E}_\lambda\left[q_\infty\right]\Big|_{\lambda=0+} = \frac{\mathbf{E}\left[\sigma^2\right]}{2\mathbf{E}\left[\sigma\right]}. \tag{4.31}$$

Recalling Proposition 4.2.4(a), which provides the second light traffic derivative, we observe that (4.31) already holds as an equality for $n = 2$, and not just in the limit as $n \to \infty$. This hints to the possibility that (4.31) holds with equality for all $n = 2, 3 \ldots$, namely

$$\frac{d^n}{d\lambda^n} \mathbf{E}_\lambda\left[q_\infty\right]\Big|_{\lambda=0+} = \frac{n!}{2} \mathbf{E}\left[\sigma\right]^{n-1} \mathbf{E}\left[\sigma^2\right], \quad n = 2, 3 \ldots. \tag{4.32}$$

If (4.32) were to hold, or were assumed to hold, then additional results would be within reach. In fact, it follows from (2.33) that (4.32) indeed holds for the expected queue size $\mathbf{E}_\lambda\left[q_\infty^{(u)}\right]$ in the system with instantaneous inputs. These observations motivate a closer examination of the relation between the discrete–time queue driven by $M|G|\infty$ inputs and the corresponding system with i.i.d. instantaneous inputs.

4.3 Gradual inputs

In this section we revisit the system with gradual input process. We aim at obtaining expressions for the expected queue length and the probability that the queue length is zero. We adapt the arguments of [55] and suppose that the system with gradual inputs operates in parallel with a second system, with instantaneous inputs. That is, we consider the Lindley recursion (2.17) with instantaneous inputs (2.16), and the sequence $\{q_n^{(a)},\ n = 0, 1, \ldots\}$ evolving according to

$$q_0^{(a)} = 0; \qquad q_{n+1}^{(a)} = [q_n^{(a)} + b_{n+1}^{(a)} - c]^+, \quad n = 0, 1, \ldots, \tag{4.33}$$

where the sequence of gradual inputs $\{b_n^{(a)},\ n = 1, 2, \ldots\}$ is given by

$$b_n^{(a)} = \sum_{k=1}^{n} \sum_{j=1}^{\beta_k} \mathbf{1}\left[\sigma_{k,j} > n - k\right], \qquad n = 0, 1, \ldots, \tag{4.34}$$

with the convention that empty sums are zero. We couple the two input sequences (4.34) and (2.16) by constructing them both from the same i.i.d. rvs $\{\beta_{n+1},\ n = 0, 1, \ldots\}$ and $\{\sigma_{n+1,i},\ n = 0, 1, \ldots,\ i = 1, 2, \ldots\}$ of Section 2.1.1.

4.3.1 Stationary version

The gradual input sequence $\{b_n^{(a)},\ n = 0, 1, \ldots\}$ of (4.34) driving recursion (4.33) is not the stationary version of the $M|G|\infty$ busy server process. This stationary version $\{b_n,\ n = 0, 1, \ldots\}$ was given by (2.1) in Section 2.1.1 and contains an additional term due to the servers initially busy. It is thus necessary to establish a relation between the stationary regimes for (4.33) (with non-stationary inputs) and for (2.11) (with stationary inputs). This is done in the following Lemma.

Lemma 4.3.1 *Under the stability condition $\lambda \mathbf{E}[\sigma] < c$ the sequences $\{q_n,\ n = 0, 1, \ldots\}$ and $\{q_n^{(a)},\ n = 0, 1, \ldots\}$ associated with the Lindley recursions (2.11) and (4.33), respectively, converge weakly to the same stationary rv q_∞.*

Proof. Define $M := \max\{\sigma_{0,j},\ j = 1, \ldots, b\}$, where b and $\{\sigma_{0,j},\ j = 1, \ldots, b\}$ are given in Assumption (A), so that M is the (finite) time that elapses until the activity periods of all initially active sessions expire. From (2.1) and (2.2) we note that

$$b_{n+M+1} = b_{n+M+1}^{(a)}, \qquad n = 0, 1, \ldots. \tag{4.35}$$

Thus, on every sample path, the queue size sequences $\{q_{n+M},\ n = 0, 1, \ldots\}$ and $\{q_{n+M}^{(a)},\ n = 0, 1, \ldots\}$ are driven by the same input sequence (4.35) with possibly

different (in general) initial conditions q_M and $q_M^{(a)}$. Because

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{j=1}^{n} b_j^{(a)} = \lambda\mathbf{E}\left[\sigma\right] \qquad a.s.$$

we can invoke Lemma 6.1.4 in [54, p. 134], to conclude that under the stability condition $\lambda\mathbf{E}\left[\sigma\right] < c$ the sequences $\{q_{n+M},\ n = 0, 1, \ldots\}$ and $\{q_{n+M}^{(a)},\ n = 0, 1, \ldots\}$ strongly couple on every sample path, hence, due to the already established fact that $q_n \Longrightarrow_n q_\infty$, weakly converge to the same stationary rv q_∞. ■

4.3.2 Case $c \leq 1$: A stochastic comparison

In this section we assume that the multiplexer release rate is $c \leq 1$, and establish a strong stochastic comparison between q_∞ and $q_\infty^{(u)}$. For each $n = 1, 2, \ldots$ let d_n denote the total amount of work processed by the server in the system (2.17) with instantaneous inputs (2.16) during the interval $(0, n]$. The sequence $\{d_n,\ n = 1, 2, \ldots\}$ satisfies

$$d_n = \sum_{k=1}^{n} u_k - q_n^{(u)}, \quad n = 1, 2, \ldots. \tag{4.36}$$

Next, look at the sibling system (4.33) with gradual inputs. Under the condition $c \leq 1$, a single active session suffices to make full use of the server capacity c, despite the gradual nature of its input. This is no longer true if $c > 1$; in that case if there is only one active session in the system it is served at unit rate and the portion $c - 1$ of the capacity remains unused. Thus, under the assumption $c \leq 1$, in the system (4.33) (with gradual inputs) the server processes, in every sample path, exactly the same amount of work as the server in the coupled system (2.16)

(with instantaneous inputs) . This crucial observation enables us to write

$$d_n = \sum_{k=1}^{n} b_k^{(a)} - q_n^{(a)}, \quad n = 1, 2, \ldots . \tag{4.37}$$

for the same completed work sequence $\{d_n,\ n = 1, 2, \ldots\}$ as that of (4.36). We can thus obtain a relation between the queue sizes in the two systems: For each $n = 1, 2, \ldots$, set

$$v_n := \sum_{k=1}^{n} \sum_{j=1}^{\beta_{n-k+1}} (\sigma_{n-k+1,j} - k)^+ \tag{4.38}$$

and note from relations (4.36) and (4.37), and definitions (4.34) and (2.16) that

$$\begin{aligned} q_n^{(u)} - q_n^{(a)} &= \sum_{k=1}^{n} \sum_{j=1}^{\beta_k} (\sigma_{k,j} - (n-k+1))^+ \\ &= \sum_{k=1}^{n} \sum_{j=1}^{\beta_{n-k+1}} (\sigma_{n-k+1,j} - k)^+ \\ &= v_n. \end{aligned} \tag{4.39}$$

Observe here that (4.39) implies the sample path inequality

$$q_n^{(a)} \leq q_n^{(u)}, \quad n = 1, 2, \ldots . \tag{4.40}$$

Take n going to infinity in (4.40) and recall that the stochastic ordering $\leq_{st}$ is stable under weak convergence (Proposition B.3 of Appendix B). It is now plain from Lemma 4.3.1 that

$$q_\infty^{(a)} \leq_{st} q_\infty^{(u)} \tag{4.41}$$

with $q_\infty^{(a)} =_{st} q_\infty$, and the following stochastic comparison is obtained:

Proposition 4.3.1 *Consider the Lindley recursions (2.11) and (2.17) with inputs characterized by a common pair (λ, σ) and release rate c such that $\lambda \mathbf{E}[\sigma] < c \leq 1$. The respective stationary rvs q_∞ and $q_\infty^{(u)}$ satisfy*

$$q_\infty \leq_{st} q_\infty^{(u)}. \tag{4.42}$$

4.3.3 Case $c \leq 1$: Expected queue size $\mathbf{E}[q_\infty]$

In addition to the stability condition $\lambda \mathbf{E}[\sigma] < c$ we assume that $\mathbf{E}\left[\sigma^2\right] < \infty$, which ensures $\mathbf{E}\left[q_\infty^{(u)}\right] < \infty$, and proceed again via (4.39) and Lemma 4.3.1 to express $\mathbf{E}[q_\infty]$ in terms of $\mathbf{E}\left[q_\infty^{(u)}\right]$. For each $n = 1, 2, \ldots$, it holds that

$$\mathbf{E}\left[q_n^{(a)}\right] = \mathbf{E}\left[q_n^{(u)}\right] - \mathbf{E}[v_n]. \tag{4.43}$$

From (4.38) by Wald's identity we have

$$\begin{aligned} \mathbf{E}[v_n] &= \lambda \sum_{k=1}^{n} \mathbf{E}\left[(\sigma - k)^+\right] \\ &= \lambda \mathbf{E}[\sigma] \sum_{k=1}^{n} \mathbf{P}\left[\widehat{\sigma} > k\right] \end{aligned}$$

whence

$$\begin{aligned} \lim_{n\to\infty} \mathbf{E}[v_n] &= \lambda \mathbf{E}[\sigma] \left(\mathbf{E}\left[\widehat{\sigma}\right] - 1\right) \\ &= \lambda \mathbf{E}[\sigma] \left(\frac{\mathbf{E}[\sigma(\sigma+1)]}{2\mathbf{E}[\sigma]} - 1\right) \\ &= \frac{\lambda}{2} \mathbf{E}[\sigma(\sigma-1)] \end{aligned} \tag{4.44}$$

where the second equality follows from (2.4). Turning to the sequence $\{q_n^{(u)},\ n = 0, 1, \ldots\}$ we note that (2.17) is driven by the i.i.d. rvs $\{u_n,\ n = 1, 2, \ldots\}$ of (2.16) and satisfies $q_0^{(u)} = 0 \leq q_1^{(u)}$, so that the monotonicity result of Proposition B.7 (Appendix B) applies. We get

$$q_n^{(u)} \leq_{st} q_\infty^{(u)}, \quad n = 1, 2, \ldots \tag{4.45}$$

and because of (4.40) we also obtain

$$q_n^{(a)} \leq_{st} q_\infty^{(u)}, \quad n = 1, 2, \ldots. \tag{4.46}$$

Since $\mathbf{E}\left[q_\infty^{(u)}\right] < \infty$ whenever $\mathbf{E}\left[\sigma^2\right] < \infty$, inequalities (4.45) and (4.46) imply that

$$\mathbf{E}\left[q_\infty^{(u)}\right] = \lim_{n\to\infty} \mathbf{E}\left[q_n^{(u)}\right] \quad \text{and} \quad \mathbf{E}\left[q_\infty^{(a)}\right] = \lim_{n\to\infty} \mathbf{E}\left[q_n^{(a)}\right] \tag{4.47}$$

by the dominated convergence theorem. Therefore, letting n go to infinity in (4.43) and using (4.44), (4.47) and Lemma 4.3.1 we collect

Proposition 4.3.2 *Consider the Lindley recursions (2.11) and (2.17) with inputs characterized by a common pair (λ, σ) and release rate c such that $\lambda\mathbf{E}[\sigma] < c \leq 1$. If $\mathbf{E}\left[\sigma^2\right] < \infty$, then the respective stationary rvs q_∞ and $q_\infty^{(u)}$ satisfy*

$$\mathbf{E}[q_\infty] = \mathbf{E}\left[q_\infty^{(u)}\right] - \frac{\lambda}{2}\mathbf{E}[\sigma(\sigma-1)]\,. \tag{4.48}$$

In the case where the multiplexer release rate is $c = 1/m$, for some integer $m = 1, 2, \ldots$, we can combine (4.48) with (2.51). Doing so yields an explicit expression for the expected queue size induced by $M|G|\infty$ inputs, namely

$$\mathbf{E}[q_\infty] = \frac{\lambda\mathbf{E}\left[\sigma^2\right](\lambda m\mathbf{E}[\sigma] + m - 1)}{2(1-\lambda m\mathbf{E}[\sigma])}, \quad m = 1, 2, \ldots, \tag{4.49}$$

provided that $\lambda\mathbf{E}[\sigma] < c = 1/m$ and $\mathbf{E}\left[\sigma^2\right] < \infty$. Of course formula (4.49) should be consistent with the light traffic derivatives of $\mathbf{E}[q_\infty]$, calculated in Section 4.2.4 through the Reiman–Simon method. Indeed, for $m = 1$, in which case $c = 1$, it is now easy to differentiate (4.49) twice and verify that (4.25) holds true.

4.3.4 Case $c = 1$: Determination of $\mathbf{P}[q_\infty = 0]$

Let the multiplexer rate be $c = 1$. We now obtain an exact expression for $\mathbf{P}[q_\infty = 0]$, i.e., the probability that the queue size is zero in the discrete–time queue with $M|G|\infty$ input process. This is accomplished by combining the corresponding result on the instantaneous inputs queue with a sample path argument for the coupled system with gradual inputs. The details are as follows:

Consider the systems (2.17) and (4.33), with coupled instantaneous and gradual inputs respectively. In Section 2.3.4 we introduced a decomposition of the queue length evolution in system (2.17) in terms of "idle" and "busy" periods $\{(I_n^{(u)}, B_n^{(u)}),\ n = 1, 2, \ldots\}$ in the sense defined by (2.40) and (2.41). In the same manner we introduce $\{(I_n, B_n),\ n = 1, 2, \ldots\}$, the idle and busy period lengths associated with (4.33). These rv pairs are also i.i.d.; let (I, B) be the corresponding generic pair. We first note that the coupled systems both start empty and have identical cycle lengths,

$$I_n + B_n = I_n^{(u)} + B_n^{(u)}, \quad n = 1, 2, \ldots, \tag{4.50}$$

so that it suffices to focus on one such common regenerative cycle, say the first one.

Next, look at system (2.17), with instantaneous inputs. Starting from $q_0^{(u)} = 0$, the queue finds itself in an idle period. In the time slot preceding the first busy period there must be either a single session arrival whose workload exceeds one, or more than one session arrivals. Now, consider the coupled system (4.33), with gradual inputs. Here, on every sample path, the queue is empty whenever it is empty in the sibling instantaneous input system. Clearly, if a busy period in (2.17) is triggered by at least two arrivals in the preceding slot, then the queue size in (4.33) also grows positive, the two systems become simultaneously busy, and the idle period I_1 for (4.33) is equal to the respective idle period $I_1^{(u)}$ for (2.17). However, if a busy period in (2.17) is initiated by a single session arrival, the queue size in (4.33) remains zero, as long as only one session is active. This is due to the assumption that, in system (4.33), an active session generates input at rate equal to the multiplexer release rate $c = 1$. In that case the idle period I_1 for (4.33) is

longer than the respective idle period $I_1^{(u)}$ for (2.17). Let

$$J := I_1 - I_1^{(u)} \tag{4.51}$$

denote their difference. Since $J = 0$ if the generic busy period $I^{(u)}$ of (2.17) starts with more than one session arrivals, it only remains to condition on the event that it starts with a single arrival.

To calculate the probability that the busy period of (2.17) is initiated by a lone arriving session we write

$$\begin{aligned} \mathbf{P}\left[I^{(u)} = k, \beta_{I^{(u)}} = 1\right] &= \mathbf{P}\left[q_1^{(u)} = \ldots = q_{k-1}^{(u)} = 0, \beta_k = 1, \sigma_{k,1} > 1\right] \\ &= \eta^{k-1}\mathbf{P}\left[\beta = 1\right]\mathbf{P}\left[\sigma > 1\right], \quad k = 1, 2, \ldots, \end{aligned}$$

with η given by (2.43), so that

$$\begin{aligned} \mathbf{P}\left[\beta_{I^{(u)}} = 1\right] &= \sum_{k=1}^{\infty}\mathbf{P}\left[I^{(u)} = k, \beta_{I^{(u)}} = 1\right] \\ &= \frac{1}{1-\eta}\mathbf{P}\left[\beta = 1\right]\mathbf{P}\left[\sigma > 1\right] \\ &= \mathbf{E}\left[I^{(u)}\right]\mathbf{P}\left[\beta = 1\right]\mathbf{P}\left[\sigma > 1\right], \end{aligned} \tag{4.52}$$

where the last equality in (4.52) follows from (2.45). Next, note that on the event $\{\beta_{I^{(u)}} = 1\}$ the queue size associated with (4.33) becomes positive only after an arrival of a second session which takes place before the activity period of the first one expires. Let the rv X count the number of successive time slots that elapse until a second session arrives, inclusive of the slot where this arrival occurs. The rv X is geometrically distributed with

$$\mathbf{P}\left[X > k\right] = \mathbf{P}\left[\beta = 0\right]^k, \quad k = 0, 1, \ldots. \tag{4.53}$$

From the discussion above it follows that the generic rv J satisfies

$$J =_{st} \begin{cases} \min([\sigma - 1 | \sigma > 1], X) & \text{if } \beta_{I^{(u)}} = 1 \\ 0 & \text{if } \beta_{I^{(u)}} \neq 1. \end{cases} \tag{4.54}$$

We can now calculate the expectation $\mathbf{E}[J]$ without difficulty. Using (4.53) we write

$$\begin{aligned}
\mathbf{E}[\min([\sigma-1|\sigma>1],X)] &= \sum_{k=0}^{\infty}\mathbf{P}[\min([\sigma-1|\sigma>1],X)>k] \\
&= \sum_{k=0}^{\infty}\mathbf{P}[\sigma-1>k|\sigma>1]\,\mathbf{P}[\beta=0]^k \\
&= \frac{1}{\mathbf{P}[\sigma>1]}\sum_{k=0}^{\infty}\mathbf{P}[\sigma>k+1]\,\mathbf{P}[\beta=0]^k \\
&= \frac{1}{\mathbf{P}[\sigma>1]\,\mathbf{P}[\beta=0]}\,\mathbf{E}[\min(\sigma,X)-1]\,.
\end{aligned}$$

Therefore, (4.54) via (4.52) implies

$$\mathbf{E}[J] = \frac{\mathbf{P}[\beta=1]}{\mathbf{P}[\beta=0]}\,\mathbf{E}\left[I^{(u)}\right]\;\mathbf{E}[\min(\sigma,X)-1]\,. \tag{4.55}$$

By application of the Renewal–Reward theorem, in conjuction with Lemma 4.3.1, we determine $\mathbf{P}[q_\infty=0]$ as

$$\mathbf{P}[q_\infty=0] = \frac{\mathbf{E}[I]}{\mathbf{E}[B]+\mathbf{E}[I]} = \frac{\mathbf{E}\left[I^{(u)}\right]+\mathbf{E}[J]}{\mathbf{E}[B^{(u)}]+\mathbf{E}[I^{(u)}]}, \tag{4.56}$$

where the second equality in (4.56) follows from (4.50) and (4.51). Thus, making use of (2.42) and (4.55) in (4.56) above we collect

Lemma 4.3.2 *Consider the Lindley recursions (2.11) and (2.17) with inputs characterized by a common pair (λ,σ), with generic $\mathbb{N}$–valued session arrival rv β such that $\lambda=\mathbf{E}[\beta]<\infty$ and release rate $c=1$. If $\lambda\mathbf{E}[\sigma]<1$, then*

$$\mathbf{P}[q_\infty=0] = \mathbf{P}\left[q_\infty^{(u)}=0\right]\left(1+\frac{\mathbf{P}[\beta=1]}{\mathbf{P}[\beta=0]}\,\mathbf{E}[\min(\sigma,X)-1]\right), \tag{4.57}$$

where the rv X follows the geometric distribution (4.53).

We stress that (4.57) is valid for any $\mathbb{N}$–valued sequence of i.i.d. rvs $\{\beta_{n+1},\ n=0,1,\ldots\}$ with $\mathbf{E}[\beta]<\infty$, as the arguments leading to Lemma 4.3.2 do not require that β be Poisson distributed.

By specializing (4.57) to $M|G|\infty$ input processes and invoking (2.31) we obtain an explicit expression for $\mathbf{P}\left[q_\infty = 0\right]$, in the case where the multiplexer release rate is $c = 1$.

Proposition 4.3.3 *Consider the Lindley recursion (2.11) with $c = 1$. Under the stability condition $\lambda\mathbf{E}\left[\sigma\right] < 1$, it holds that*

$$\mathbf{P}\left[q_\infty = 0\right] = (1 - \lambda\mathbf{E}\left[\sigma\right])e^\lambda \left(1 + \lambda \sum_{k=1}^{\infty} e^{-\lambda k}\mathbf{P}\left[\sigma > k\right]\right). \tag{4.58}$$

4.3.5 Case $c = 1$: Short– vs long–range dependence

Let the multiplexer release rate be $c = 1$. In this section we seek to develop some understanding as to what kind of light traffic results should be expected when Assumption (C) fails, as would be the case if σ follows some subexponential distribution. To that end we discuss expansions based on the closed form expression (4.58) for $\mathbf{P}\left[q_\infty = 0\right]$. For short–range dependent inputs we verify that (4.58) is in agreement with the partial light traffic information for $\mathbf{P}\left[q_\infty = 0\right]$ obtained in Section 4.2.4. More interestingly, we show that (4.58) can be exploited to provide a light traffic limit for long–range dependent inputs; such a result could not have been obtained via the Reiman–Simon theory (at least in its present form).

We start by considering the case where the $M|G|\infty$ input process is short–range dependent. From (4.58) it follows that

$$\mathbf{P}\left[q_\infty > 0\right] = 1 - (1 - \lambda\mathbf{E}\left[\sigma\right])e^\lambda \left(1 + \lambda \sum_{k=1}^{\infty} (e^{-\lambda k} - 1)\mathbf{P}\left[\sigma > k\right] + \lambda(\mathbf{E}\left[\sigma\right] - 1)\right). \tag{4.59}$$

We now show that the leading term in the light traffic expansion of $\mathbf{P}\left[q_\infty > 0\right]$ is of order λ^2 and evaluate this term explicitly. To do this note that the mapping $k \to \frac{1}{\lambda}(1 - e^{-\lambda k})$ monotonically increases to the mapping $k \to k$ as $\lambda \to 0+$.

Therefore, since $\mathbf{E}\left[\sigma^2\right] < \infty$ under short-range dependence, we conclude that

$$\lim_{\lambda \to 0+} \sum_{k=1}^{\infty} \frac{1}{\lambda} \left(1 - e^{-\lambda k}\right) \mathbf{P}\left[\sigma > k\right] = \sum_{k=1}^{\infty} k \, \mathbf{P}\left[\sigma > k\right] = \frac{1}{2} \, \mathbf{E}\left[\sigma(\sigma - 1)\right] \tag{4.60}$$

by the monotone convergence theorem, while

$$\lim_{\lambda \to 0+} \frac{1}{\lambda^2} \left(1 - e^{\lambda}(1 - \lambda \mathbf{E}\left[\sigma\right])(1 - \lambda + \lambda \mathbf{E}\left[\sigma\right])\right) = \mathbf{E}\left[\sigma\right] \left(\mathbf{E}\left[\sigma\right] - 1\right) + \frac{1}{2}. \tag{4.61}$$

Thus, combining (4.61) and (4.60) with (4.59) yields

Corollary 4.3.1 (Short–range dependence) *In the setup of Proposition 4.3.3, with* $\mathbf{E}\left[\sigma^2\right] < \infty$*, it holds that*

$$\lim_{\lambda \to 0+} \frac{1}{\lambda^2} \, \mathbf{P}\left[q_\infty > 0\right] = \frac{1}{2} \, \left(\mathbf{E}\left[\sigma(\sigma - 1)\right] + 2\mathbf{E}\left[\sigma\right] \left(\mathbf{E}\left[\sigma\right] - 1\right) + 1\right). \tag{4.62}$$

This is precisely the result implied by (4.16) and (4.27) for $b = 0$. In addition, we observe that Assumption (C) is superfluous in this case, for what is needed to obtain (4.62) is that $\mathbf{E}\left[\sigma^2\right]$ be finite. This indicates that the conclusions of Section 4.2.4 may still be valid when σ is subexponentially distributed with $\mathbf{E}\left[\sigma^2\right] < \infty$.

Next, we turn our attention to the situation where the tail of σ is regularly varying with index $-\alpha$ $(1 < \alpha < 2)$, i.e., of the form

$$\mathbf{P}\left[\sigma > n\right] = n^{-\alpha} h(n), \quad n = 1, 2, \ldots \tag{4.63}$$

for some slowly varying function $h : \mathbb{R}_+ \to \mathbb{R}_+$. In this case $\mathbf{E}\left[\sigma^2\right] = \infty$, the associated $M|G|\infty$ process is long–range dependent and the arguments used above do not apply. First write

$$\sum_{k=1}^{\infty} e^{-\lambda k} \mathbf{P}\left[\sigma > k\right] = \frac{1}{e^{-\lambda} - 1} \left(\sum_{k=1}^{\infty} e^{-\lambda k} \mathbf{P}\left[\sigma = k\right] - e^{-\lambda}\right) \tag{4.64}$$

and manipulate (4.58) to obtain

$$\begin{aligned} \mathbf{P}\left[q_\infty > 0\right] \;\; = \;\; & 1 - e^{\lambda}(1 - \lambda \mathbf{E}\left[\sigma\right]) \left(1 - \lambda - \frac{\lambda^2}{e^{-\lambda} - 1} \mathbf{E}\left[\sigma\right]\right) \\ & - \frac{\lambda e^{\lambda}}{e^{-\lambda} - 1} \, (1 - \lambda \mathbf{E}\left[\sigma\right]) \left(\sum_{k=1}^{\infty} e^{-\lambda k} \mathbf{P}\left[\sigma = k\right] - 1 + \lambda \mathbf{E}\left[\sigma\right]\right). \end{aligned} \tag{4.65}$$

Expression (4.65) has the advantage of explicitly displaying the Laplace–Stieltjes transform of the distribution of σ. This puts us in position to invoke a Tauberian result on the asymptotic behavior of Laplace–Stieltjes transforms at the origin. In particular, Theorem 8.1.6 in [7, p. 333] provides the asymptotics of the second term in (4.65) as

$$\sum_{k=1}^{\infty} e^{-\lambda k}\mathbf{P}\left[\sigma = k\right] - 1 + \lambda\mathbf{E}\left[\sigma\right] \sim \lambda^{\alpha} h(1/\lambda)\, \frac{\Gamma(2-\alpha)}{\alpha - 1} \qquad (\lambda \to 0+). \tag{4.66}$$

For the first term in (4.65), making use of the fact

$$\lim_{\lambda \to 0} \frac{e^{-\lambda} - 1 + \lambda}{\lambda^{\alpha}} = 0, \quad 1 < \alpha < 2$$

we find after straightforward calculations that

$$\lim_{\lambda \to 0} \frac{1}{\lambda^{\alpha}} \left(1 - e^{\lambda}(1 - \lambda\mathbf{E}\left[\sigma\right]) \left(1 - \lambda - \frac{\lambda^2}{e^{-\lambda} - 1}\mathbf{E}\left[\sigma\right]\right)\right) = 0. \tag{4.67}$$

Consequently, (4.65) via (4.66) and (4.67) leads to

Corollary 4.3.2 (Long–range dependence) *In the setup of Proposition 4.3.3, with the tail of σ given by (4.63), it holds that*

$$\lim_{\lambda \to 0+} \frac{\lambda^{-\alpha}}{h(1/\lambda)}\, \mathbf{P}\left[q_{\infty} > 0\right] = \frac{\Gamma(2-\alpha)}{\alpha - 1}. \tag{4.68}$$

The effect of the distribution of σ, and in particular of its second moment, on the light traffic asymptotics of $\mathbf{P}\left[q_{\infty} > 0\right]$ is now more apparent: As $\lambda \to 0+$, the "busy" queue probability $\mathbf{P}\left[q_{\infty} > 0\right]$ exhibits a λ^2 decay under short–range dependence, and the slower non-integer power decay λ^{α}, when σ satisfies (4.63), in which case the $M|G|\infty$ process is long–range dependent. The limit (4.68) prompts us to conjecture that, when $c = 1$, under long–range dependence, such a λ^{α} decay in λ of the tail probability $\mathbf{P}\left[q_{\infty} > b\right]$ holds more generally for all $b \geq 0$. Yet this problem remains as of now unresolved.

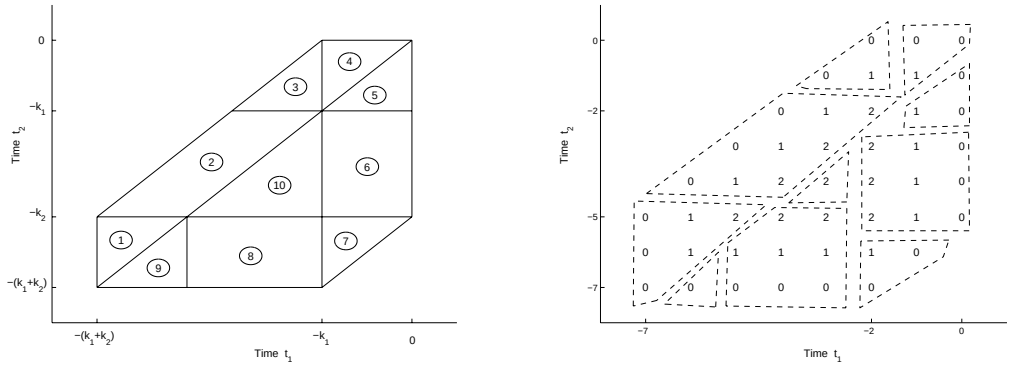

Figure 4.4: Calculation of $\widetilde{\psi}(k_1, k_2)$; example for $k_1 = 2$, $k_2 = 5$.

4.4 Proof of Lemma 4.2.1

We consider parts (a), (b) and (c) separately.

Proof of (a). By symmetry it suffices to determine $\widetilde{\psi}(k_1, k_2)$ for $k_1 < k_2$. To do this, we refer to Figure 4.3, and split the area where $\psi(\{t_1, t_2; k_1, k_2\}) > 0$ in ten numbered regions, adopting the convention that edges between region i and any other higher numbered region all belong to region i. The splitting we choose is shown in Figure 4.4, accompanied by a simple example for the case $k_1 = 2$, $k_2 = 5$.

The double sum (4.19) is broken up into sums over each region, which are calculated below.

Region 1: $$\sum_{t_1=-(k_1+k_2)}^{-k_2} \sum_{t_2=t_1}^{-k_2} (t_1 + k_1 + k_2) = \frac{k_1(k_1+1)(k_1+2)}{6}$$

Region 2: $$\sum_{t_1=1-(k_1+k_2)}^{1-k_2} (t_1 - (1-k_2) + k_1)(k_2 - k_1) = \frac{k_1(k_2-k_1)(k_1+1)}{2}$$

Region 3: $$\sum_{t_1=-2k_1+1}^{-k_1} \sum_{t_2=-k_1+1}^{t_1+k_1} (t_1 - t_2 + k_1) = \frac{k_1(k_1+1)(k_1-1)}{6}$$

Region 4: $$\sum_{t_1=1-k_1}^{0} \sum_{t_2=t_1}^{0} (-t_2) = \frac{k_1(k_1+1)(k_1-1)}{6}$$

Region 5: $\displaystyle\sum_{t_1=1-k_1}^{0}\sum_{t_2=-k_1}^{t_1-1}(-t_1)=\frac{k_1(k_1-1)(k_1+1)}{6}$

Region 6: $\displaystyle\sum_{t_1=-k_1}^{0}\sum_{t_2=-k_2}^{-k_1-1}(-t_1)=\frac{k_1(k_2-k_1)(k_1+1)}{2}$

Region 7: $\displaystyle\sum_{t_1=-k_1}^{1}\sum_{t_2=t_1-k_2}^{-k_2-1}(t_2-t_1+k_2)=\frac{k_1(k_1-1)(k_1+1)}{6}$

Region 8: $\displaystyle -k_1+\sum_{t_2=-(k_1+k_2)}^{-k_2}(t_2+k_1+k_2)(k_2-k_1)=\frac{k_1((k_1+1)(k_2-k_1)-2)}{2}$

Region 9: $\displaystyle\sum_{t_1=-(k_1+k_2)+1}^{-k_2}\sum_{t_2=-(k_1+k_2)}^{t_1-1}(t_2+k_1+k_2)-\sum_{t_2=-(k_1+k_2)}^{-k_2-1}(t_2+k_1+k_2)$

$$=\frac{k_1(k_1-1)(k_1-2)}{6}$$

Region 10: $\displaystyle\sum_{t_1=-k_2+1}^{-k_1}\sum_{t_2=-k_2}^{t_1-1}k_1-(2k_2-2k_1-1)k_1$

$$=\frac{k_1(k_2-k_1-1)(k_2-k_1-2)}{2}$$

Adding the results from regions 1–10 shows that the claim holds true. ■

Proof of (b). It suffices to evaluate $\widetilde{\psi}$ for $k_1<k_2$. We refer to the proof of Part (a), split the t_1t_2–plane in the same manner and carry out the algebra for the double sum (4.19) corresponding to $\psi:=q_0^2$. The summands are listed below.

Region 1: $\displaystyle\sum_{t_1=-(k_1+k_2)}^{-k_2}\sum_{t_2=t_1}^{-k_2}(t_1+k_1+k_2)^2=\frac{k_1(k_1+1)^2(k_1+2)}{12}$

Region 2: $\displaystyle(k_2-k_1)\sum_{t_1=1-(k_1+k_2)}^{1-k_2}(t_1-(1-k_2)+k_1)^2$

$$=\frac{k_1(k_1+1)(2k_1+1)(k_2-k_1)}{6}$$

Region 3: $\displaystyle\sum_{t_1=-2k_1+1}^{-k_1}\sum_{t_2=-k_1+1}^{t_1+k_1}(t_1-t_2+k_1)^2=\frac{k_1^2(k_1^2-1)}{12}$

Region 4: $\displaystyle\sum_{t_1=1-k_1}^{0}\sum_{t_2=t_1}^{0}(-t_2)^2=\frac{k_1^2(k_1^2-1)}{12}$

Region 5: $\displaystyle\sum_{t_1=1-k_1}^{0}\sum_{t_2=-k_1}^{t_1-1}(-t_1)^2=\frac{k_1^2(k_1^2-1)}{12}$

Region 6: $\displaystyle\sum_{t_1=-k_1}^{0}\sum_{t_2=-k_2}^{-k_1-1}(-t_1)^2=\frac{k_1(k_1+1)(2k_1+1)(k_2-k_1)}{6}$

Region 7: $\displaystyle\sum_{t_1=-k_1}^{1}\sum_{t_2=t_1-k_2}^{-k_2-1}(t_2-t_1+k_2)^2=\frac{k_1^2(k_1^2-1)}{12}$

Region 8: $\displaystyle(k_2-k_1)\sum_{t_2=-(k_1+k_2)}^{-k_2}(t_2+k_1+k_2)^2-k_1^2$

$$=\frac{k_1((k_1+1)(2k_1+1)(k_2-k_1)-6k_1)}{6}$$

Region 9: $\displaystyle\sum_{t_1=-(k_1+k_2)+1}^{-k_2}\sum_{t_2=-(k_1+k_2)}^{t_1-1}(t_2+k_1+k_2)^2-\sum_{t_2=-(k_1+k_2)}^{-k_2-1}(t_2+k_1+k_2)^2$

$$=\frac{k_1(k_1-1)^2(k_1-2)}{12}$$

Region 10: $\displaystyle\sum_{t_1=-k_2+1}^{-k_1}\sum_{t_2=-k_2}^{t_1-1}k_1^2-(2k_2-2k_1-1)k_1^2$

$$=\frac{k_1^2(k_1-k_2+1)(k_1-k_2+2)}{2}$$

Adding the summands from regions 1–10 verifies the asserted expression. ■

Proof of (c). By symmetry we need only calculate $\widetilde{\psi}$ for $k_1<k_2$. We go back to Figure 4.3, where $\widetilde{\psi}$ corresponding to $\psi:=q_0$ was displayed. Clearly, in order for q_0 to be greater than b it is necessary that $k_1>b$. If this is not true, i.e. if $k_1\le b$, then $\mathbf{1}\,[q_0>b]$ is zero, hence $\widetilde{\psi}$ corresponding to $\psi:=\mathbf{1}\,[q_0>b]$ is also zero and

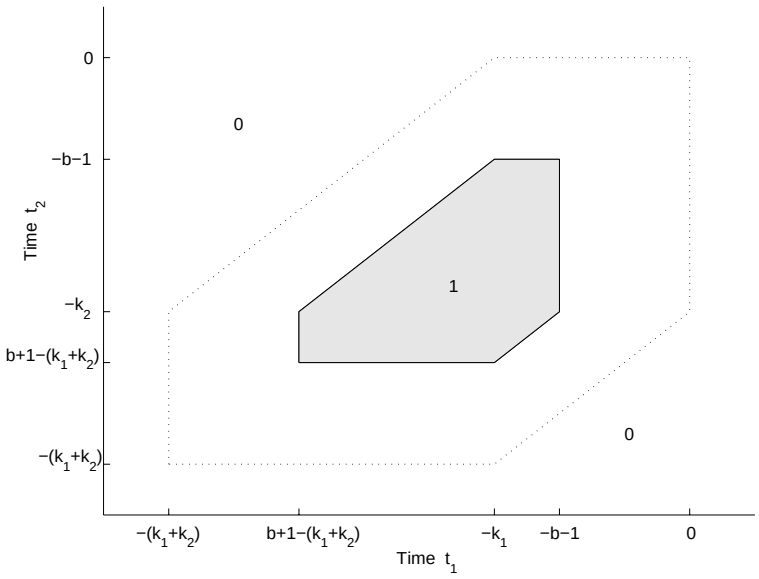

Figure 4.5: Values of $\mathbf{1}\left[q_0 > b\right](\{t_1, t_2; k_1, k_2\})$ for $k_1 < k_2$.

the second part of the claim holds true. Now, if $k_1 > b$ then the t_1t_2–plane does indeed contain a region where $\psi = 1$. We show $\psi(\{t_1, t_2; k_1, k_2\})$ corresponding to $\psi := \mathbf{1}\left[q > b\right]$ and identify this region in Figure 4.5. Using this graph, calculation of $\widetilde{\psi}(k_1, k_2)$ from the double sum (4.6) is a matter of algebra. From the points along the line $t_1 = -k_1$, which splits the $\psi = 1$ region in two trapezoids, we collect

$$\sum_{i=-(k_1+k_2)+b+1}^{-b-1} 1 = k_1 + k_2 - 2b - 1. \tag{4.69}$$

On the right trapezoid the double sum can be calculated as

$$\sum_{j=0}^{k_2-b-2} (j + k_1 - b) = \frac{1}{2}(2k_1 + k_2 - 3b - 2)(k_2 - b - 1). \tag{4.70}$$

On the left trapezoid we find

$$\sum_{j=0}^{k_1-b-2} (k_1 + k_2 - 2b - 2 - j) = \frac{1}{2}(2k_2 + k_1 - 3b - 2)(k_1 - b - 1). \tag{4.71}$$

Adding (4.69), (4.70), (4.71) and manipulating validates the claim for $k_1 > b$. ■

Chapter 5

Interpolation approximations

5.1 Introduction

In this chapter we address the problem of evaluating the buffer content distribution at a multiplexer fed by an $M|G|\infty$ arrival process, across the entire range of system utilizations between zero and one. For arbitrary session duration distribution G the system lacks the desired Markovian structure and obtaining exact analytical results appears very difficult, if not impossible. To circumvent the difficulties of an exact analysis one typically seeks to devise approximations, by relying on information gleaned from asymptotic regimes. A promising approach consists of deriving approximations from the analysis of large buffer asymptotics [29, 39, 47]; these estimates are exact in the limit as the buffer level goes to infinity. Here, we propose an alternative class of simple approximations, justified by the characterizations of the buffer content distribution in the heavy and light traffic regimes, obtained in Chapters 3 and 4, respectively. These approximations are termed interpolation approximations [49], because they arise by suitably interpolating between the heavy and light traffic limits of the quantity of interest. Such approximations are asymptotically exact in the limits as the system utilization goes to zero and

one. We provide examples of interpolation approximations to the buffer content distribution, for several commonly chosen G, covering both short– and long–range dependent $M|G|\infty$ inputs. Comparisons with simulation estimates suggest that the approximants capture accurately the queue size distribution at small buffer sizes, for which approximations based on large buffer asymptotics are often ill fitted. On the other hand, when G has finite exponential moment, we do not expect the heavy–light traffic interpolation to be accurate for buffer sizes much larger than the maximum burst length: The approximation exhibits the appropriate exponential decay in the buffer size, yet the rate is only asymptotically exact as the system utilization tends to one, i.e., in the heavy traffic limit. This drawback is often absent under long–range dependence, as there are cases where the queue size distribution has hyperbolic (in the buffer size) asymptotics with the same exponent for all traffic intensities! Moreover, an approximation is more valuable in the presence of heavy tails, when considering that alternative estimates by means of simulation take an unreasonably long time to obtain.

5.2 Summary of asymptotics

In the context of the Lindley recursion (2.11), describing the evolution of the buffer content at a multiplexer with $M|G|\infty$ inputs, we are interested in approximating the probability that the stationary buffer content exceeds b when the system utilization is $\rho := \lambda \mathbf{E}[\sigma]/c$,

$$P(b,\rho) := \mathbf{P}_\lambda [q_\infty > b], \quad b \geq 0, \quad 0 \leq \rho < 1. \tag{5.1}$$

Let us point out that the system dynamics depend on λ, G and c jointly, and not simply on the utilization ρ. However, for developing approximate expressions it is

convenient to fix G and c and adopt the view suggested by (5.1), that is consider the buffer content distribution as a function of the system utilization ρ. In a similar manner we introduce the moments of the stationary buffer content

$$Q_k(\rho) := \mathbf{E}_\lambda \left[{q_\infty}^k \right], \quad 0 \leq \rho < 1, \quad k = 1, 2, \ldots;$$

these should not be confused with the heavy traffic queue length process $\{Q(t),\ t \geq 0\}$ of Chapter 3.

The interpolation approximations we have in mind hinge on the availability of explicit expressions for limits of system quantities as $\lambda \to c/\mathbf{E}[\sigma]$ (heavy traffic limits), and derivatives with respect to λ as $\lambda \to 0$ (light traffic derivatives). For notational convenience we now rephrase the required light and heavy traffic asymptotics in terms of the utilization ρ, with the understanding that when $\rho \to 0$ or $\rho \to 1$ it is actually λ which goes to the corresponding limit, while both c and σ remain fixed.

We start from the results in the light traffic regime, obtained by the Reiman–Simon method:

Proposition 5.2.1 *Consider the Lindley recursion (4.1) with release rate $c \geq 1$ and let $b \geq 0$. If Assumption (C) is satisfied, then the following hold:*

(a) *For each $n = 0, 1, \ldots, \lfloor c \rfloor$, we have*

$$\frac{d^n}{d\rho^n} P(b, 0+) = 0; \tag{5.2}$$

(b) *In addition, for $c = 1$, we have*

$$\begin{aligned} \mathbf{E}[\sigma]^2 \frac{d^2}{d\rho^2} P(b, 0+) &= \mathbf{E}\left[(\sigma - b)^{+2} \right] \mathbf{P}[\sigma > b] + 2\, \mathbf{E}\left[(\sigma - b)^+ \right]^2 \\ &\quad -3\, \mathbf{E}\left[(\sigma - b)^+ \right] \mathbf{P}[\sigma > b] + \mathbf{P}[\sigma > b]^2 . \end{aligned} \tag{5.3}$$

Considered next is the behavior of the queue with $M|G|\infty$ arrivals in heavy traffic, that is, as the packet arrival rate $\lambda\mathbf{E}[\sigma]$ tends to the multiplexer release rate c from below. We tacitly assume that the heavy traffic limit of the stationary distribution coincides with the stationary distribution of the heavy traffic limit. Using Assumption (B) we express the relevant facts from Theorems 3.4.1 and 3.4.3 in terms of the utilization ρ in the following

Proposition 5.2.2 *The heavy traffic limits of the stationary queue length distribution associated with (2.11) can be classified as follows:*

(a) *If* $\mathbf{E}\left[\sigma^2\right] < \infty$, *then*

$$\lim_{\rho\to 1}\mathbf{P}_\lambda\left[(1-\rho)\ q_\infty > x\right] = \exp\left(-\frac{2\mathbf{E}[\sigma]}{\mathbf{E}\left[\sigma^2\right]}\ x\right), \quad x \geq 0. \tag{5.4}$$

(b) *If* $\mathbf{P}[\sigma > n] = n^{-\alpha}$, $n = 1, 2, \ldots$, *with* $1 < \alpha < 2$, *then*

$$\lim_{\rho\to 1}\mathbf{P}_\lambda\left[(1-\rho)^{1/(\alpha-1)} q_\infty > x\right] = E_{\alpha-1}\left(-\frac{(\alpha-1)\mathbf{E}[\sigma]}{\Gamma(2-\alpha)}\ x^{\alpha-1}\right), \quad x \geq 0, \tag{5.5}$$

where for $\nu > 0$, $E_\nu : \mathbb{R} \to \mathbb{R}$ *is the Mittag–Leffler special function defined by (3.19).*

Part (a) of Proposition 5.2.2 addresses the classical short–range dependent case, for which the heavy traffic normalizer is $1-\rho$ and the limiting heavy traffic distribution is exponential. Part (b) deals with a long–range dependent $M|G|\infty$ arrival process, in which case the heavy traffic queue length distribution is expressed through a Mittag–Leffler function with hyperbolic decay, while the heavy traffic normalizer is $(1-\rho)^{1/(\alpha-1)}$ and has power–law behavior.

The results under the light and heavy traffic regimes are subsequently combined into approximations for all values of ρ in the interval $[0, 1)$.

5.3 Heavy–light traffic interpolations

Whenever Assumption (C) is satisfied, $\mathbf{P}_\lambda[q_\infty > b]$ is infinitely differentiable with respect to ρ at $\rho = 0$, hence it can be approximated by bringing together heavy traffic limits and light traffic derivatives into a Taylor series–like expansion. To this end we enforce Assumption (C) throughout Sections 5.3.1 and 5.3.2 and follow the approach proposed in [21]. In passing, we also discuss approximations for $Q_k(\rho)$, $k = 1, 2$. The details are given below:

5.3.1 Tail probability approximations

Consider the normalized queue length rv $(1-\rho)\, q_\infty$ and define

$$\begin{aligned} F(x,\rho) &:= \mathbf{P}_\lambda[(1-\rho)\, q_\infty > x] \\ &= P\left(\frac{x}{1-\rho},\ \rho\right), \qquad 0 \le \rho < 1, \quad x \ge 0 \end{aligned} \tag{5.6}$$

and

$$F(x,1) := \lim_{\rho\to 1} \mathbf{P}_\lambda[(1-\rho)\, q_\infty > x], \quad x \ge 0. \tag{5.7}$$

Assume that partial derivatives of $F(x,\rho)$ with respect to ρ, up to order n, at $\rho = 0+$, are available. Construct $\widehat{F}_n(x,\rho)$, the n^{th} order interpolation approximation to $F(x,\rho)$, by means of the polynomial

$$\widehat{F}_n(x,\rho) := \sum_{i=0}^{n} \frac{\rho^i}{i!} \frac{\partial^i}{\partial \rho^i}\ F(x,0+) + \left(F(x,1) - \sum_{i=0}^{n} \frac{1}{i!} \frac{\partial^i}{\partial \rho^i}\ F(x,0+)\right) \rho^{n+1}. \tag{5.8}$$

Observe that

$$\widehat{F}_n(x,1) = F(x,1) \quad \text{and} \quad \frac{\partial^i}{\partial \rho^i}\ \widehat{F}_n(x,0+) = \frac{\partial^i}{\partial \rho^i}\ F(x,0+), \quad i = 0,1,\ldots,n,$$

that is, $\widehat{F}_n(x,\rho)$ is precisely that unique $n+1$ degree polynomial in ρ which matches the $n+1$ partial derivatives of $F(x,\rho)$ at $\rho = 0+$ and its heavy traffic limit.

Now, by reversing the $(1-\rho)$ normalization in $\widehat{F}_n(x,\rho)$ we generate the n^{th} order interpolation approximation to $\mathbf{P}_\lambda[q_\infty > b]$ as

$$\mathbf{P}_\lambda[q_\infty > b] \approx \widehat{F}_n((1-\rho)\ b, \rho). \tag{5.9}$$

Note that, in principle, this may lie outside $[0,1]$, in which case it is obviously a poor approximation.

To calculate the quantities associated with (5.9) it remains to express the partial derivatives appearing in (5.8) in terms of the light traffic derivatives of $\mathbf{P}_\lambda[q_\infty > b]$. We have

$$\frac{\partial}{\partial\rho}\ F(x,0+) = \frac{\partial}{\partial\rho}\ P(x,0+) + x\ \frac{\partial}{\partial x}P(x,0+) \tag{5.10}$$

and

$$\begin{aligned}\frac{\partial^2}{\partial\rho^2}\ F(x,0+) &= \frac{\partial^2}{\partial\rho^2}\ P(x,0+) + 2x\ \frac{\partial^2}{\partial\rho\partial x}P(x,0+) \\ &\quad +2x\ \frac{\partial}{\partial x}P(x,0+) + x^2\ \frac{\partial^2}{\partial x^2}P(x,0+).\end{aligned} \tag{5.11}$$

In case additional light traffic information is available, repeated application of the chain rule will yield higher order derivatives, as needed.

We are now ready to write approximate expressions anchored on the heavy and light traffic results of Section 5.2. Proposition 5.2.2(a) provides the limit (5.7) that should be inserted in (5.8). Proposition 5.2.1(a) can be used to substitute for the partials in (5.10) and then in (5.8). Thus, if the multiplexer release rate is $c \geq 1$, the $\lfloor c \rfloor^{th}$ order interpolation approximation to $\mathbf{P}_\lambda[q_\infty > b]$ is simply

$$\mathbf{P}_\lambda[q_\infty > b] \approx \widehat{F}_{\lfloor c \rfloor}((1-\rho)\ b, \rho) = \rho^{\lfloor c \rfloor + 1} \exp\left(-\frac{2\mathbf{E}[\sigma]}{\mathbf{E}[\sigma^2]}\ (1-\rho)\ b\right). \tag{5.12}$$

More can be accomplished in the case $c = 1$, since Proposition 5.2.1(b) affords us a promising 2^{nd} order interpolation approximation. Insertion of (5.11) in (5.8) yields

$$\widehat{F}_2(b,\rho) = \frac{1}{2}\rho^2(1-\rho)\ \frac{\partial^2}{\partial\rho^2}\ P(b,0+) + \rho^3 \exp\left(-\frac{2\mathbf{E}[\sigma]}{\mathbf{E}[\sigma^2]}\ b\right) \tag{5.13}$$

and the latter leads to the 2^{nd} order approximation

$$\mathbf{P}_\lambda\left[q_\infty > b\right] \approx \widehat{F}_2((1-\rho)\ b, \rho), \quad c = 1, \tag{5.14}$$

where Proposition 5.2.1(b) is used to supply the second partial derivative in (5.13).

5.3.2 Moment approximations

Next, we briefly deal with moment approximations. We restrict attention to the case $c = 1$ and consider only the queue length first and second moment. The relevant light traffic limits are given by (4.15), (4.25) and (4.26). In heavy traffic we see from (5.4) that

$$\lim_{\rho\to 1}(1-\rho)^k\ Q_k(\rho) = k!\left(\frac{\mathbf{E}\left[\sigma^2\right]}{2\mathbf{E}\left[\sigma\right]}\right)^k, \quad k = 1, 2, \ldots.$$

Moment approximations are then developed by interpolating for $(1-\rho)Q_1(\rho)$ and $(1-\rho)^2\ Q_2(\rho)$, in very much the same manner as distribution approximations. We skip the details of the derivation and list the resulting final expressions

$$Q_1(\rho) = \frac{\mathbf{E}\left[\sigma^2\right]}{2\mathbf{E}\left[\sigma\right]}\ \frac{\rho^2}{1-\rho}, \qquad c = 1 \tag{5.15}$$

and

$$Q_2(\rho) \approx \frac{\rho^2}{4(1-\rho)^2}\left(\rho\left(\frac{\mathbf{E}\left[\sigma^2\right]^2}{\mathbf{E}\left[\sigma\right]^2} - 1\right) + \frac{\mathbf{E}\left[\sigma^2\right]^2}{\mathbf{E}\left[\sigma\right]^2} + 1\right), \quad c = 1. \tag{5.16}$$

Note that formula (5.15) is in fact exact, as it coincides with result (4.49) (with $m = 1$) and with a continuous time analog established for a fluid model in [55, p. 23]. This match clearly validates the interpolation method. On the contrary we note that (5.16) cannot be exact. To see this, consider the example where $\sigma = 1$ deterministic. This corresponds to i.i.d. Poisson arrivals, for which a probability

generating function of the queue length rv is available. Using (2.32) it can be shown that the exact expression is

$$Q_2(\rho) = \frac{\rho^2}{6(1-\rho)^2}\ (\rho^2 - \rho + 3), \quad c = 1, \ \ \sigma = 1 \ \ a.s.$$

a formula that clearly cannot be recovered using only two light traffic derivatives. Still, when $\sigma = 1$ approximation (5.16) is within 9% of the correct value, for all ρ in $[0, 1)$.

5.3.3 Long–range dependence

It is apparent from the developments of Section 5.2 that the light traffic results, as stated in Proposition 5.2.1, do not cover several interesting distributions belonging to the subexponential family. Such is for example the lognormal distribution, which violates Assumption (C) despite having finite k^{th} moment for all $k = 0, 1, \ldots$. In view of Corollary 4.3.1, it is natural to expect that Assumption (C) can be relaxed to require that $\mathbf{E}\left[\sigma^k\right]$ be finite, for appropriate $k \geq 2$, in order for Proposition (5.2.1) to go through. This would still not address the case of long–range dependence, characterized by $\mathbf{E}\left[\sigma^2\right] = \infty$. We are however able to construct a sharp approximation, based on heuristic arguments presented below.

Consider the Pareto distribution $\mathbf{P}\left[\sigma > n\right] = n^{-\alpha}$, $n = 1, 2, \ldots$, with $1 < \alpha < 2$; in this case not only Assumption (C) fails, but as seen from Proposition 5.2.1, (5.3) yields infinity. This indicates that $\mathbf{P}_\lambda\left[q_\infty > b\right]$ may not be an analytic function of ρ under long–range dependence. When $c = 1$ we recall that Corollary 4.3.2 strongly suggests that $\lim_{\rho \to 0} \rho^{-\alpha} \mathbf{P}_\lambda\left[q_\infty > b\right]$ is the sought after non-trivial limit, for all $b \geq 0$. These considerations lead us to postulate that, when $c = 1$,

$$\lim_{\rho \to 0} \rho^{-\alpha} \mathbf{P}_\lambda\left[q_\infty > b\right] = K(b) \tag{5.17}$$

for some unknown mapping $K : \mathbb{R}_+ \to \mathbb{R}_+$, which, by Corollary 4.3.2, satisfies $K(0) = \Gamma(2-\alpha)/(\alpha-1)\mathbf{E}[\sigma]^\alpha$. On the other hand, the heavy traffic result of Proposition 5.2.2(b) hints at developing an approximation around the normalized rv $(1-\rho)^{1/(\alpha-1)}\ q_\infty$. Then, taking advantage of Proposition 5.2.2(b) we propose the approximant

$$\mathbf{P}_\lambda[q_\infty > b] \approx E_{\alpha-1}\left(-\frac{(\alpha-1)\mathbf{E}[\sigma]}{\Gamma(2-\alpha)}\ \frac{1-\rho}{\rho^\alpha}\ b^{\alpha-1}\right), \quad c=1. \tag{5.18}$$

This expression is in agreement with the heavy traffic limit (5.5). In addition, from the Mittag–Leffler function asymptotics given in [17, p. 207], we have

$$E_{\alpha-1}(-x) \sim \frac{1}{x}\frac{1}{\Gamma(2-\alpha)} \quad (x \to \infty)$$

which ensures that, as $\rho \to 0$, approximation (5.18) conforms with the conjectured light traffic limit (5.17).

We close the presentation of the approximate expresssions with a comment. Recall that each active source in the $M|G|\infty$ arrival process generates one information unit per time slot. So, $c = 1$ corresponds to the case where the amount of service in one slot is exactly equal to the amount of information that one active source generates in one slot. When $c = 1$ a single active source suffices to make full use of the server capacity; in this system there is never any leftover capacity to simultaneously serve more than one source. On the contrary, when $c > 1$, the server can attend to more than one source during one time slot, so that there is a multiple service feature to the system behavior. An exact or approximate analysis in this regime is clearly more challenging.

5.4 Numerical results

To gauge the accuracy of the proposed expressions we have carried out simulation experiments under various choices for the distribution of the session duration rv σ. The experimental values are obtained by regenerative simulation and relative widths accompanying them correspond to 95% confidence intervals. We almost exclusively (with one exception) confine ourselves to the simple situation where the multiplexer release rate is $c = 1$. While the list of examples below is not exhaustive, it does serve to illustrate the ability of the heavy–light traffic interpolation to "ballpark" the true tail probabilities, as well as its limitations.

Deterministic When the session duration is deterministic, $\sigma = D$ *a.s.*, for some positive integer D, approximation (5.14) reads

$$\begin{aligned}\mathbf{P}_\lambda\left[q_\infty > b\right] \;\approx\; & \frac{\rho^2(1-\rho)}{2D^2}\left(3[D-(1-\rho)b]^+([D-(1-\rho)b]^+ - 1)\right. \\ & \left. + \mathbf{1}\left[D > (1-\rho)b\right]\right) + \rho^3 \exp\left(-\frac{2}{D}(1-\rho)b\right), \qquad b = 0, 1, \ldots .\end{aligned}$$

We let the session duration be $\sigma = 3$ and obtain simulation estimates for the steady state probability $\mathbf{P}_\lambda\left[q_\infty > 0\right]$. Of course, in this case the exact expression (4.58) is available. In Table 5.1 we list simulation estimates and numerical values from (4.58) and from the light–heavy traffic interpolation. A comparison of the exact formula (4.58) to the light–heavy traffic interpolation shows that, in this case, the agreement is excellent. Since we expect the approximation to be asymptotically exact at the endpoints $\rho = 0$ and $\rho = 1$, it is not surprising that the largest errors occur in moderate traffic.

In the same setup, we next consider the tail probability $\mathbf{P}_\lambda\left[q_\infty > 4\right]$. No exact expressions are available in this case. From Table 5.2 we see that although the

	Tail probability $\mathbf{P}_\lambda[q_\infty > 0]$			
ρ	Exact	Simulation	Approximation	Error (%)
0.1	1.0478e-02	1.0469e-02±0.2%	1.0500e-02	–0.21
0.2	4.1622e-02	4.1668e-02±0.3%	4.1778e-02	–0.38
0.3	9.3042e-02	9.3020e-02±0.2%	9.3500e-02	–0.49
0.4	1.6441e-01	1.6442e-01±0.2%	1.6533e-01	–0.56
0.5	2.5545e-01	2.5533e-01±0.2%	2.5694e-01	–0.58
0.6	3.6594e-01	3.6607e-01±0.1%	3.6800e-01	–0.56
0.7	4.9573e-01	4.9601e-01±0.1%	4.9817e-01	–0.49
0.8	6.4470e-01	6.4488e-01±0.1%	6.4711e-01	–0.37
0.9	8.1279e-01	8.1264e-01±0.1%	8.1450e-01	–0.21

Table 5.1: $\mathbf{P}_\lambda[q_\infty > 0]$ for deterministic session duration $\sigma = 3$.

approximation yields estimates in the correct order of magnitude, the errors are substantial when not in the moderate–to–heavy traffic regime. This can be explained as follows: When $\sigma = 3$, in order for the queue to build up to 4 at least 3 sources should be simultaneously active. Note that the light traffic component of the approximation consists of the second derivative, which can be obtained by considering sample paths with at most two source activations in the system. Thus, any effects due to the activation of more than two sources are not adequately accounted for in light traffic.

Uniform We now specialize (5.14) to the case where σ is uniformly distributed on $\{1, \ldots, M\}$, i.e., $\mathbf{P}[\sigma = n] = 1/M, \;\; n = 1, 2, \ldots, M$. This yields

$$\mathbf{P}_\lambda[q_\infty > b] \;\approx\; \mathbf{1}[M > (1-\rho)b] \frac{\rho^2(1-\rho)}{3M^2(M+1)^2} \left(1 + 5(M - (1-\rho)b)^2\right)$$

	Tail probability $\mathbf{P}_\lambda[q_\infty > 4]$		
ρ	Simulation	Approximation	Error (%)
0.1	1.1271e-04±1.7%	9.0718e-05	19.51
0.2	1.3444e-03±1.7%	9.4753e-04	29.52
0.3	6.1736e-03±0.9%	5.9952e-03	2.89
0.4	1.9246e-02±0.6%	1.4415e-02	25.10
0.5	4.7745e-02±0.5%	3.9894e-02	16.44
0.6	1.0292e-01±0.4%	9.5777e-02	6.94
0.7	2.0035e-01±0.3%	1.9757e-01	1.39
0.8	3.6093e-01±0.3%	3.6379e-01	–0.79
0.9	6.1407e-01±0.3%	6.1902e-01	–0.81

Table 5.2: $\mathbf{P}_\lambda[q_\infty > 4]$ for deterministic session duration $\sigma = 3$.

$$\times(M-(1-\rho)b)^2+\rho^3\exp\left(-6\,\frac{(1-\rho)b}{2M+1}\right), \qquad b=0,1,\ldots.$$

For $M = 5$ we compare simulation vs approximation in Tables 5.3 and 5.4, for system utilizations $\rho = 0.2$ and $\rho = 0.8$ respectively. Once more, the approximation is very sharp for small buffer sizes. As the buffer size increases beyond the maximum burst length and the true probabilities become smaller, the approximation lingers on in the correct order of magnitude, but clearly deteriorates away from heavy traffic. Eventually, as the buffer size tends to infinity, the interpolation approximation overestimates the actual probabilities.

	Tail probability $\mathbf{P}_\lambda[q_\infty > b]$		
Buffer size	Simulation	Approximation	Error (%)
0	4.4907e-02±0.2%	4.5333e-02	–0.95
2	1.1747e-02±0.5%	1.1399e-02	2.97
5	1.4543e-03±1.4%	9.7380e-04	33.04
8	1.9456e-04±3.7%	2.4379e-04	–25.30
10	5.1620e-05±7.0%	1.0186e-04	–97.31

Table 5.3: Utilization $\rho = 0.2$; $\sigma \sim$ uniform$(1, 5)$.

	Tail probability $\mathbf{P}_\lambda[q_\infty > b]$		
Buffer size	Simulation	Approximation	Error (%)
0	6.5489e-01±0.1%	6.6133e-01	–0.98
10	2.0086e-01±0.4%	1.9161e-01	4.60
20	6.3303e-02±0.9%	5.8057e-02	8.29
30	1.9964e-02±1.7%	1.9406e-02	2.79
40	6.2827e-03±3.1%	6.5188e-03	–3.75
50	1.9703e-03±5.6%	2.1897e-03	–11.13

Table 5.4: Utilization $\rho = 0.8$; $\sigma \sim$ uniform$(1, 5)$.

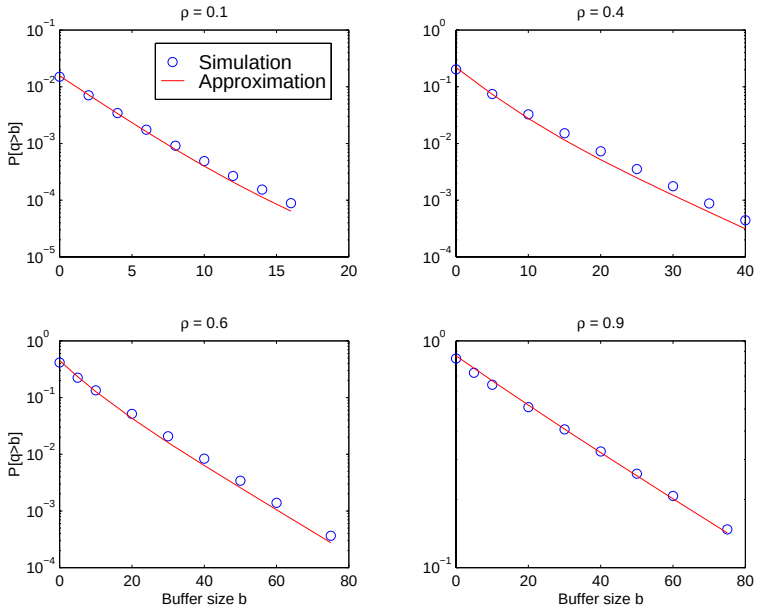

Figure 5.1: Geometric $\gamma = 0.8$ session duration.

Geometric Taking σ to follow the geometric distribution, $\mathbf{P}[\sigma > n] = \gamma^n$, $n = 0, 1, \ldots$, with $0 < \gamma < 1$ we obtain from (5.14) that

$$\mathbf{P}_\lambda[q_\infty > b] \approx \frac{\rho^2}{2}(1-\rho)(1+\gamma)^2\gamma^{2(1-\rho)b} + \rho^3 \exp\left(-2\,\frac{1-\gamma}{1+\gamma}(1-\rho)b\right), \quad b = 0, 1, \ldots$$

As an example we set $\gamma = 0.8$ and plot simulated and approximate values in Figure 5.1, for system utilizations $\rho = 0.1$, 0.4, 0.6 and 0.9. In all cases confidence interval widths were within 10% of the mean. The linear decrease of the simulated values suggests an exponential decay of the queue length distribution, in agreement with large deviations results. Figure 5.1 clearly indicates that the heavy–light traffic interpolation is sufficient for providing rough estimates for a wide range of probabilities and buffer sizes.

An example for $c > 1$ If the multiplexer release rate is $c > 1$, expression (5.14) is not applicable, the first non-zero light traffic derivative is not available, and the

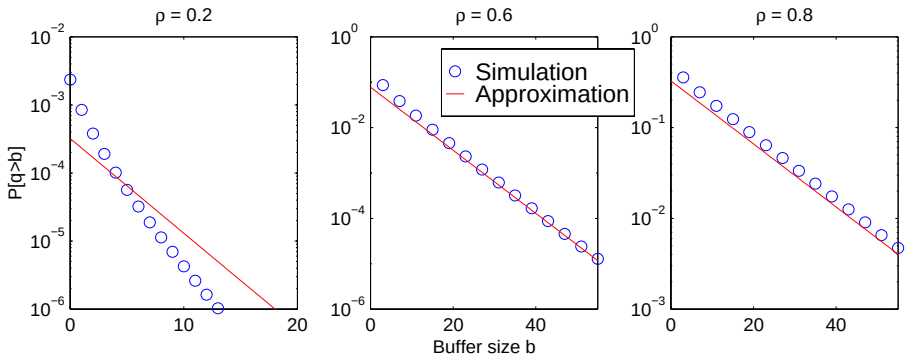

Figure 5.2: Geometric $\gamma = 2/3$ session duration; release rate $c = 4$.

only available heavy–light traffic approximant is (5.12). To illustrate its behavior we provide an example in Figure 5.2, where we have picked $c = 4$ and σ geometric with parameter $\gamma = 2/3$. The results correspond to system utilizations $\rho = 0.2$, $\rho = 0.6$ and $\rho = 0.8$. The approximation fares very well in moderate to heavy loads, yet it obviously yields inaccurate results for $\rho = 0.2$ due to insufficient light traffic information.

Pareto Let the session duration rv follow the Pareto distribution $\mathbf{P}[\sigma > n] = n^{-\alpha}$, $n = 1, 2, \ldots$, with $1 < \alpha < 2$, which case the $M|G|\infty$ process is long–range dependent and the approximate expression (5.18) is in effect. Assessing the performance of (5.18) requires numerical evaluation of the Mittag–Leffler function. In general, a calculation based on the series expansion (3.19) is not recommended. Instead, the Laplace transform of the Mittag–Leffler law can be inverted by contour integration along a suitably chosen path in the complex plane; details are deferred to Section 5.5. We finally arrive at the alternative expression

$$E_\nu(-x) = \frac{\sin(\nu\pi)}{\nu\pi} \int_0^{\pi/2} \frac{e^{-(x\tan\theta)^{1/\nu}}}{1 + \sin(2\theta)\cos(\nu\pi)}\, d\theta, \qquad x \geq 0, \quad 0 < \nu < 1, \quad (5.19)$$

which is evaluated by numerical integration.

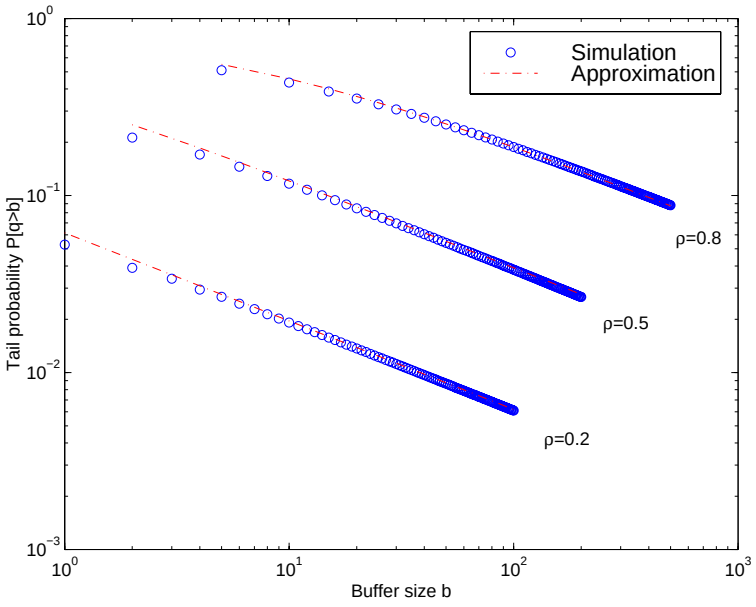

Figure 5.3: Pareto $\alpha = 1.5$ session duration.

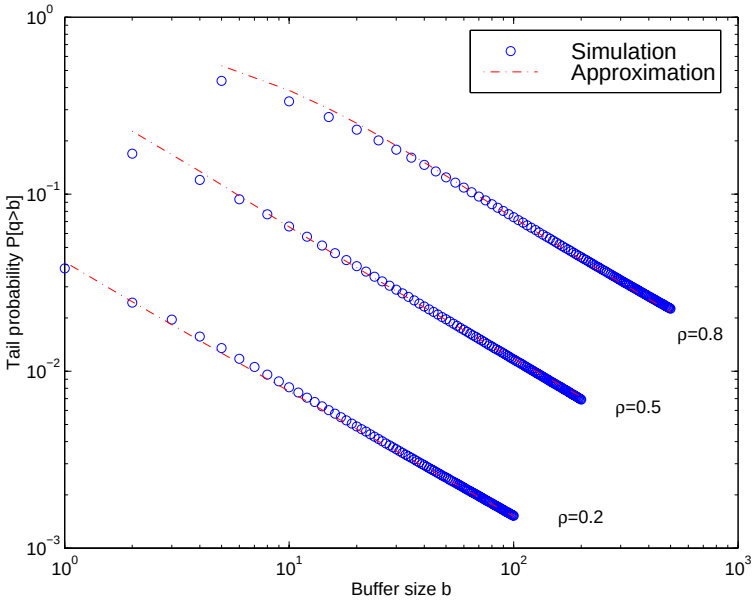

Figure 5.4: Pareto $\alpha = 1.7$ session duration.

We have tested approximation (5.18) for system utilizations $\rho = 0.2$, 0.5 and 0.8. Under long–range dependence simulation estimates converge very slowly; moreover confidence intervals based on the regenerative method cannot be constructed, because the underlying period has infinite variance. In the results shown the runs were 10^9 time slots long, and by that time the estimates had stabilized. The log–log scale plots in Figures 5.3 and 5.4 correspond to two Pareto distributions with parameters $\alpha = 1.5$ and $\alpha = 1.7$. Observe that the heavier $\alpha = 1.5$ Pareto tail induces larger tail probabilities than $\alpha = 1.7$, at the same system utilizations. In both Figures 5.3 and 5.4 we see that simulated and approximate values are very close, suggesting that expression (5.18) provides a satisfactory approximation. Note also the almost linear shape of the curves in log–log scale, reflecting the power law asymptotics of the queue size distribution announced in [29, 39, 46].

Truncated Pareto When σ follows a truncated Pareto distribution, the resulting $M|G|\infty$ process is short–range dependent. Yet, over a finite range of time scales, it can display dependencies similar to those of a long–range dependent $M|G|\infty$ process. Specifically, for $1 < \alpha < 2$, pick some $N = 2, 3, \ldots$ and consider the truncated Pareto distribution on $\{1, 2, \ldots, N\}$ given by

$$\mathbf{P}\left[\sigma > n\right] = \frac{1}{1 - N^{-\alpha}} \left(n^{-\alpha} - N^{-\alpha}\right), \quad n = 1, 2, \ldots, N. \tag{5.20}$$

The distribution (5.20) has finite support and clearly satisfies Assumption (C). The integer parameter N provides the desired flexibility in controlling the tail behavior of σ, hence the dependencies in the $M|G|\infty$ arrivals. When $N = 2$ the corresponding rv σ is deterministic, $\sigma = 2$ *a.s.* As N increases the second moment of the session duration distribution also increases, thus leading to stronger dependencies in the $M|G|\infty$ process. In the limit as N goes to infinity the rv σ

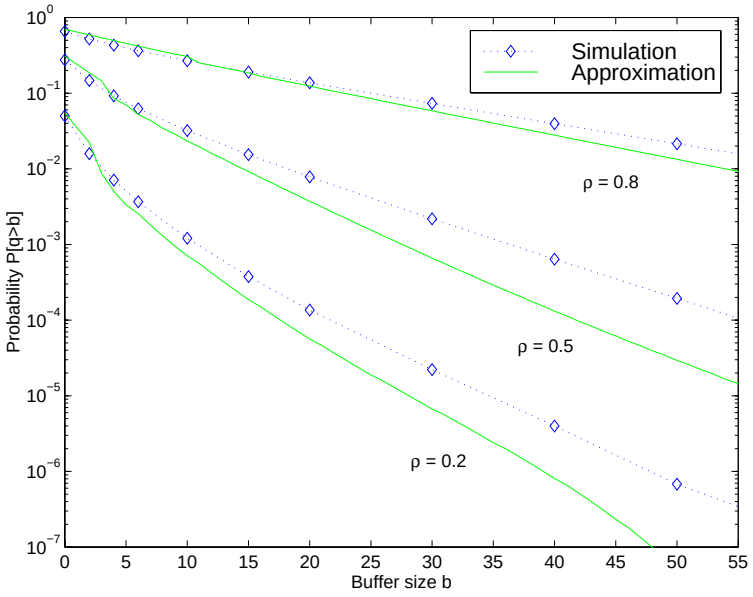

Figure 5.5: Truncated Pareto $\alpha = 1.7$, $N = 50$.

converges weakly to the standard Pareto rv

$$\mathbf{P}\left[\sigma > n\right] = n^{-\alpha}, \quad n = 1, 2, \ldots,$$

so that we traverse the boundary from a rv σ with finite exponential moment to one that has infinite variance, hence to a corresponding $M|G|\infty$ process that is long–range dependent.

To study the effect of the truncation level N on the queue size distribution we carry out simulation experiments for $\alpha = 1.7$ and two truncation levels $N = 50$ and $N = 1000$. Results are collected for system utilizations $\rho = 0.2, 0.5, 0.8$. As approximation (5.14) does not assume any simple closed form expression in this case, we calculate the various quantities entering (5.14) numerically.

In Figure 5.5 we compare simulation results for truncation level $N = 50$ with the approximate values obtained from (5.14). Confidence intervals widths are not shown, since, with the exception of the three points at the bottom of the

plot, they were well within 10% of the mean. The pairs of curves, corresponding to $\rho = 0.2, 0.5$ and 0.8, show that the approximation tracks the true queue size probabilities satisfactorily, especially for small buffer sizes. However, it is clear that as the target probabilities become smaller and the buffer size of interest larger, the quality of the approximation degrades.

Further increase in N is expected to be even more revealing of the limitations of approximation (5.14). Note that, maintaining $\alpha = 1.7$ and increasing N from 50 to 1000 results in a small increase in the expectation, from $\mathbf{E}[\sigma] = 2.9$ to $\mathbf{E}[\sigma] = 3.035$, and a large increase in the second moment, from $\mathbf{E}\left[\sigma^2\right] = 15.863$ to $\mathbf{E}\left[\sigma^2\right] = 42.49$. Figure 5.6 depicts the queue size probabilities for truncated Pareto session duration with $\alpha = 1.7$ and $N = 1000$, at system utilizations $\rho = 0.2, 0.5$ and 0.8. Along with simulation estimates we also plot the values obtained from approximation (5.14) (labeled SRD) and those from expression (5.18) (labeled LRD). The latter is appropriate for a long–range dependent $M|G|\infty$ process, so it does not strictly apply to the truncated Pareto setup. However (5.18) becomes applicable in the limit as the truncation level N goes to infinity. Thus, it is representative of the shape to which the simulation curve tends as N grows larger. From Figure 5.6 it becomes clear that while the estimates from the second order interpolation (5.14) are adequate for small buffer sizes, they fail to track the true probabilities as the buffer size increases. In fact, for $N = 1000$ and within the buffer range shown in the figures, it is the curve from the LRD approximation (5.18) that is closer to the simulated values.

Obviously, the larger the variance of the truncated Pareto rv, the closer the simulation curve will be to that of the LRD approximation (5.18). Moreover, we see that for a fixed variance of σ the match between the simulation curve and that

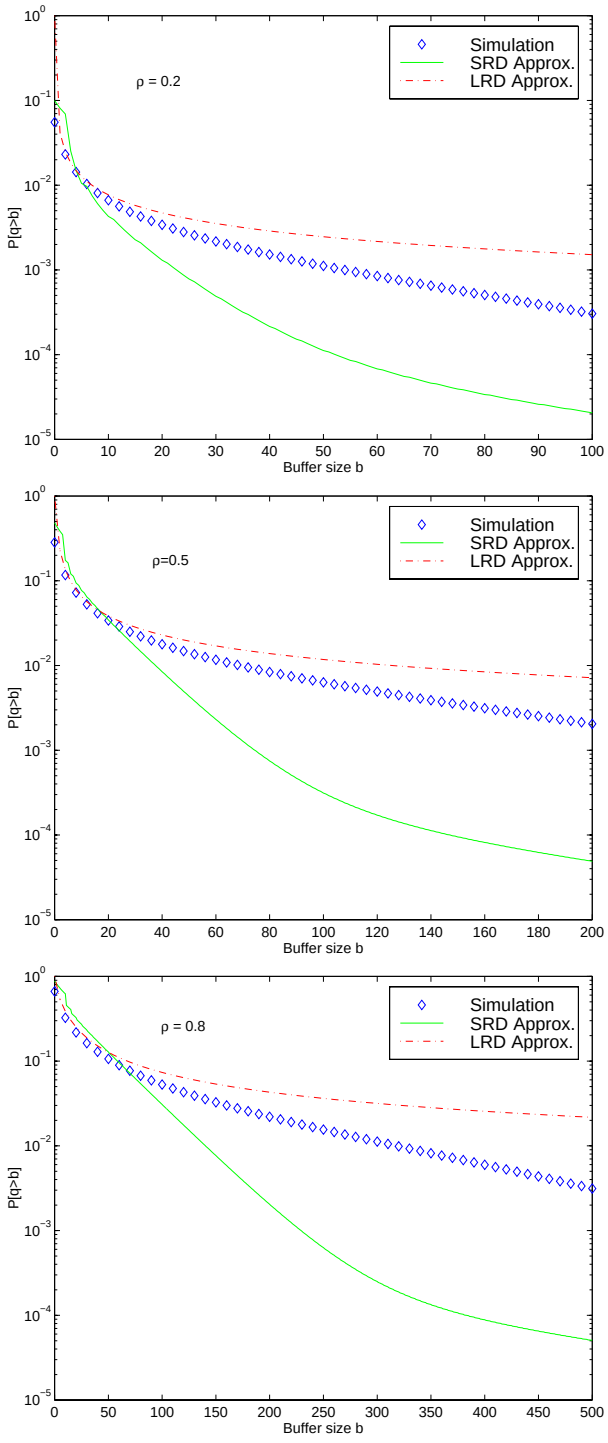

Figure 5.6: SRD vs LRD approximation for truncated Pareto session duration.

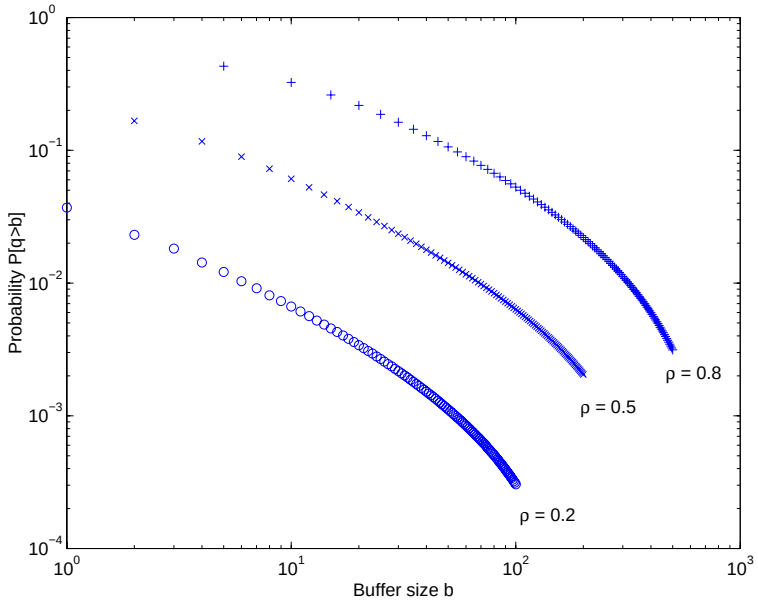

Figure 5.7: Truncated Pareto $\alpha = 1.7$, $N = 1000$.

of the LRD approximation is better at smaller buffer sizes. On the other hand, as the buffer size grows to infinity, the LRD approximation (5.18) will also eventually fail: It exhibits a hyperbolic decay, while, as follows from the developments in [47], the queue size distribution induced by truncated Pareto session durations decays exponentially fast. This can also be verified visually, from the linear shape of the rightmost part of the simulation curves in Figure 5.6. This contrast between the truncated Pareto and standard Pareto session durations is more evident when comparing the log–log plots of simulated values in Figures 5.4 and 5.7. Although the shape of the left part of the curves in Figure 5.7 is almost linear, suggestive of a hyperbolic initial segment, the righmost part decreases rapidly and the curves become concave, in accordance with the anticipated exponential decay.

5.5 On the Mittag–Leffler distribution

In Chapter 3 we saw that the heavy traffic buffer content distributions for long–range dependent $M|G|\infty$ arrivals are given through Mittag–Leffler special functions. Here, we leave the series expansion (3.19) for the Mittag–Leffler function aside, and discuss instead an alternative representation. This offers an interesting interpretation for this class of distributions and is often more amenable to numerical calculations, such as the ones performed in Section 5.4.

It is known [17, p. 207] that for $0 < \nu < 1$, the Mittag–Leffler function $E_\nu(-x)$ is completely monotone on $[0, \infty)$, i.e., for all $n = 0, 1, \ldots$

$$(-1)^n \frac{d^n}{dx^n} E_\nu(-x) \geq 0, \qquad x \geq 0, \quad 0 < \nu < 1.$$

The class of completely monotone functions is characterized by the following theorem [20, p. 439], due to Bernstein:

Theorem 5.5.1 *The function $\phi : [0, \infty) \to \mathbb{R}$ is completely monotone if and only if it is of the form*

$$\phi(\lambda) = \int_0^{+\infty} e^{-\lambda x} \, F(dx) \tag{5.21}$$

where F is a measure, not necessarily finite, on $[0, \infty)$.

It is then clear that for $0 < \nu < 1$, the Mittag–Leffler function $x \to E_\nu(-x)$, $x \geq 0$, admits the representation (5.21) and hence can be viewed as a mixture of exponential distributions. Its corresponding measure F in (5.21) is determined as follows:

First, we recall the established Laplace transform relations. In particular, (3.20) and (3.23) provide the Laplace transform pair

$$\int_0^{+\infty} e^{-sx} E_\nu(-x^\nu) \, dx = \frac{1}{s} \, \frac{1}{1 + s^{-\nu}}, \qquad s \geq 0, \quad 0 < \nu < 1; \tag{5.22}$$

this formula can also be derived from [17, Eq. (18) p. 209]. We obtain a representation of the form (5.21) for $E_\nu(-x^\nu)$ by using the definition of the Laplace inversion integral

$$E_\nu(-x^\nu) = \lim_{Y\to+\infty} \frac{1}{2\pi j} \int_{d-jY}^{d+jY} e^{sx}\, \frac{1}{s}\, \frac{1}{1+s^{-\nu}}\, ds, \qquad x \geq 0, \quad 0 < \nu < 1, \quad (5.23)$$

where we pick d to be any strictly positive abscissa on the real axis. The denominator $1 + s^{-\nu}$ has roots of the form $s_m = e^{-j\frac{2m+1}{\nu}\pi}$, $m = 0, \pm 1, \ldots$, and since $0 < \nu < 1$ we have $\dfrac{|2m+1|}{\nu} > 1$ for every $m = 0, \pm 1, \ldots$, so that there is no root with argument in $[-\pi, \pi]$. Thus, the integral (5.23) can be evaluated along the clockwise path from H to A shown in Figure 5.8, namely

$$E_\nu(-x^\nu) = \lim_{\substack{R\to+\infty \\ r\to 0}} \frac{1}{2\pi j} \int_{H}^{A} e^{sx}\, \frac{1}{s}\, \frac{1}{1+s^{-\nu}}\, ds, \qquad x \geq 0, \quad 0 < \nu < 1. \quad (5.24)$$

In the limit as the radius R goes to infinity the contribution of the arcs CB and GF to the integral above is zero, by Jordan's lemma. The limiting contribution of the arcs HG and BA is also zero, because their length is bounded and e^{sx} is also bounded along these arcs. Finally, on the circular arc ED we set $s = r\, e^{j\phi}$, ϕ in $[-\pi, \pi]$, and see that the resulting integrand vanishes as $r \to 0$. Thus, in (5.24) only the integrals along the segments FE and DC remain in the limit. Setting $s = ye^{-j\pi}$ and $s = ye^{j\pi}$, $y > 0$, for FE and DC, respectively, and manipulating we collect

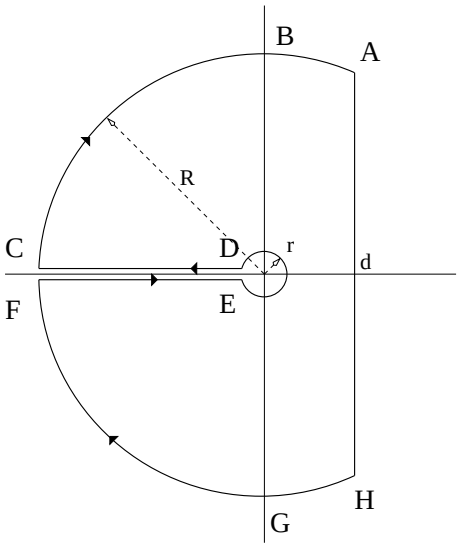

Figure 5.8: Integration path

$$E_\nu(-x^\nu) = \int_0^{+\infty} e^{-xy}\, f_\nu(y)\, dy, \qquad x \geq 0, \quad 0 < \nu < 1, \quad (5.25)$$

with

$$f_\nu(y) := \frac{\sin(\nu\pi)}{\pi y^{1-\nu}\,(y^{2\nu} + 2y^\nu \cos(\nu\pi) + 1)}, \qquad y \geq 0, \quad 0 < \nu < 1. \tag{5.26}$$

Expression (5.19) is now obtained by a change of variable $\tan\theta = y^\nu$ in (5.25). Relations (5.25) and (5.26) provide the desired interpretation of the Mittag–Leffler distribution in terms of an infinite mixture of exponentials. The specific form of (5.26) shows that when $0 < \nu < 1$, the density function $f_\nu(y)$ weighting an exponential distribution with parameter y increases to infinity as y goes to zero, so that the Mittag–Leffler distribution contains no single dominant exponential with strictly positive parameter, as expected. This infinite mixture of exponentials is to be contrasted with the classical heavy traffic queue size distribution under short–range dependence, given by a single exponential; the one to which expression (5.25) collapses as $\nu \to 1$.

Representation (5.25) and convergence to the exponential distribution as $\nu \to 1$ are illustrated in Figure 5.9. The bottom plots show the density function $f_\nu(x)$ for $\nu = 0.6$, 0.8 and 0.99, and the top log–log scale plots show the corresponding Mittag–Leffler distributions $E_\nu(-x^\nu)$. The dash–dotted line depicts the negative exponential e^{-x}. It is clear that as $\nu \to 1$ the density $f_\nu(x)$ tends to place all of its mass at one, i.e., at a single exponential, and this also becomes apparent from the top plot for $\nu = 0.99$. In addition, we observe that, even for $\nu = 0.99$, although e^{-x} and $E_\nu(-x^\nu)$ are very close for small values of the argument x, they remain strikingly different as x grows to infinity.

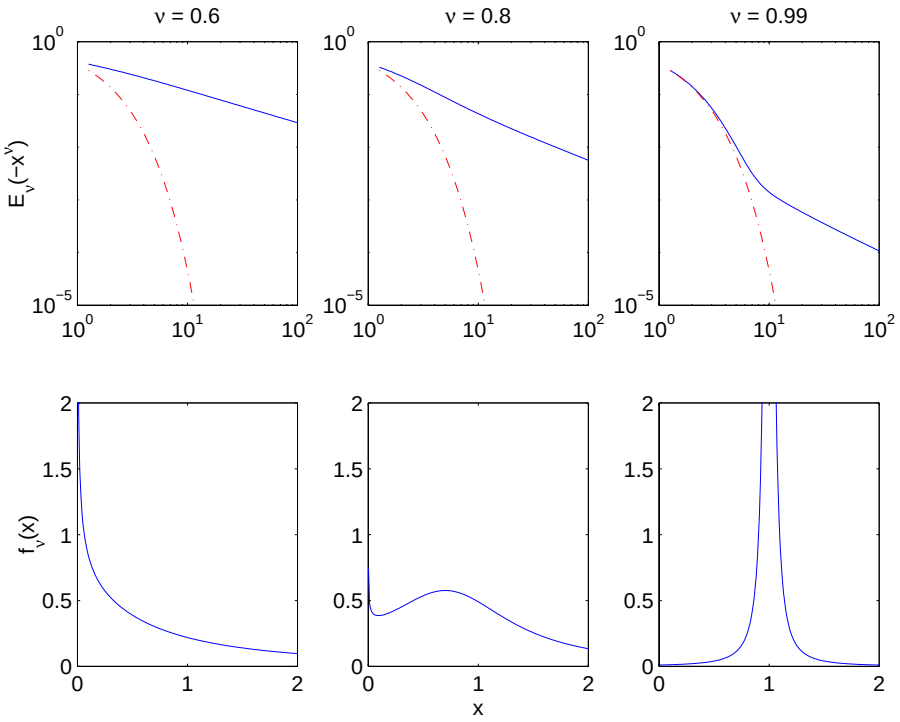

Figure 5.9: Mittag–Leffler distributions $E_\nu(-x^\nu)$ and spectral densities $f_\nu(x)$.

Chapter 6

Conclusions

Recent measurements studies have demonstrated that communication networks carry traffic much burstier than expected (self–similar, long–range, or subexponentially dependent). These findings have generated interest in the potential implications of high variability and dependence on network performance.

In this dissertation, we sought to understand the impact of (strong) correlations in the input packet stream on the performance of a single network multiplexer. This was modeled as a discrete–time queue driven by a family of $M|G|\infty$ correlated arrival processes. Given that obtaining exact solutions is, in general, extremely difficult, we instead focused on the analysis of the system behavior in two asymptotic regimes, namely light and heavy traffic.

In heavy traffic, we distinguished between $M|G|\infty$ arrival processes with short– and long–range dependence, identifying for each case the appropriate heavy traffic scaling that results in non–degenerate limits. The resulting limits for short–range dependent inputs involve the standard Brownian motion. Of particular interest are the conclusions for the long–range dependent case: The normalized queue length can be expressed as a function not of a fractional Brownian motion, but of an α–stable, $1/\alpha$ self–similar independent increments Lévy process. The buffer content

distribution in heavy traffic is expressed through a Mittag–Leffler special function and displays a hyperbolic decay, of power $1 - \alpha$.

Investigation of the system behavior in light traffic reveals the effect of two aspects of the $M|G|\infty$ arrivals, i.e., the session duration distribution G and the gradual nature of the inputs, as opposed to the instantaneous inputs of a standard $GI|GI|1$ queue. However, the arising limits cannot be fully classified in terms of short vs long–range dependence property of the $M|G|\infty$ process, hence demonstrating that the latter is not the only factor that impacts performance.

Exploiting the results above, we proposed a family of heuristic approximations for a multiplexer with $M|G|\infty$ inputs. These interpolation approximations were developed by combining the asymptotic characterizations of the buffer content distribution under heavy and light traffic conditions and are applicable to all traffic intensities. For several common pmfs G the approximants assume a simple final form, and are capable of providing quick and reliable estimates of the buffer content distribution, especially for small buffer sizes.

In closing, we mention two open questions. First, Corollary 4.3.2 lead us to conjecture that, in light traffic, if $c = 1$ and the session duration distribution has a regularly varying tail of order $-\alpha$, with $1 < \alpha < 2$, then the buffer content distribution also exhibits a power law behavior of the form λ^{α}. Second, the stochastic comparisons discussed here could be complemented by stochastic comparisons within the family of $M|G|\infty$ inputs. For two discrete–time queues, with $c = 1$, each driven by instantaneous inputs (2.16), representation (2.38) shows that a convex comparison between the respective distributions of σ translates into a strong stochastic comparison between the corresponding queue lengths. This can be viewed as a discrete–time analog of the well known result for the wait-

ing time in a standard $M|G|1$ queue and is another instance of a folk theorem of queueing "variability increases delays". We conjecture that the folk theorem holds for the gradual $M|G|\infty$ inputs as well (when $c = 1$), and simulation results indeed suggest that a convex stochastic ordering between session durations leads to a strong stochastic comparison between the queue lengths. However, as the analog to (2.38) is not available, the conjecture remains unsettled, and may require multidimensional stochastic comparisons between $M|G|\infty$ arrival vectors.

Appendix A

Asymptotic invertibility of regularly varying functions

We establish the asymptoptic invertibility property given in Proposition 3.3.1 by relying on the following facts:

Proposition A.1 *Consider a Borel measurable function* $\beta : \mathbb{R}_+ \to \mathbb{R}$ *such that* $\lim_{x\to\infty} \beta(x) = 0$. *With* $A > 0$ *and* $y > 0$, *the equation*

$$y = \exp\left(\int_A^x \frac{1-\beta(t)}{t}\, dt\right), \quad x \geq A \tag{A.1}$$

has a unique solution $x := x(y)$ *for all* y *large enough. Moreover,*

$$\lim_{y\to\infty} \frac{x(\gamma y)}{x(y)} = \gamma, \quad \gamma > 0, \tag{A.2}$$

or equivalently, the mapping $y \to x(y)$ *is regularly varying of order* 1, *i.e.,* $x(y) \sim yv(y)$ $(y \to \infty)$ *for some slowly varying function* $v : \mathbb{R}_+ \to \mathbb{R}_+$.

Proof. Set

$$B(x) := \int_A^x \frac{1-\beta(t)}{t}\, dt, \quad x \geq A$$

and pick ε in $(0,1)$. Since $\lim_{x\to\infty}\beta(x)=0$, there exists $A^\star=A^\star(\varepsilon)>A$ such that

$$\frac{1-\varepsilon}{t}<\frac{1-\beta(t)}{t}<\frac{1+\varepsilon}{t},\quad t\geq A^\star. \tag{A.3}$$

It is straightforward to see that $\lim_{x\to\infty}B(x)=\infty$ and $B(A)=0$, and by continuity, the range of $x\to\exp(B(x))$ contains the semi–infinite interval $[1,\infty)$. We also conclude from (A.3) that $x\to B(x)$ is strictly monotone increasing on the interval $[A^\star,\infty)$, and the existence and uniqueness of a solution to (A.1) follows whenever $y\geq y^\star$ with $y^\star:=\exp(B(A^\star))$. The solution mapping $y\to x(y)$ is strictly increasing on $[y^\star,\infty)$.

We now turn to proving (A.2). There is nothing to prove when $\gamma=1$. With $\gamma>1$, whenever $y\geq y^\star$, we get

$$\gamma=\frac{\gamma y}{y}=\exp(B(x(\gamma y))-B(x(y)))=\exp\left(\int_{x(y)}^{x(\gamma y)}\frac{1-\beta(t)}{t}\,dt\right),$$

and the use of the inequalities (A.3) yields

$$\left[\frac{x(\gamma y)}{x(y)}\right]^{1-\varepsilon}\leq\gamma\leq\left[\frac{x(\gamma y)}{x(y)}\right]^{1+\varepsilon},$$

or equivalently,

$$\gamma^{\frac{1}{1+\varepsilon}}\leq\frac{x(\gamma y)}{x(y)}\leq\gamma^{\frac{1}{1-\varepsilon}}. \tag{A.4}$$

Letting y go to infinity in (A.4), we conclude

$$\gamma^{\frac{1}{1+\varepsilon}}\leq\liminf_{y\to\infty}\frac{x(\gamma y)}{x(y)}\leq\limsup_{y\to\infty}\frac{x(\gamma y)}{x(y)}\leq\gamma^{\frac{1}{1-\varepsilon}},$$

and (A.2) is obtained as we note that ε is arbitrary in $(0,1)$. The case $\gamma<1$ is handled in a similar way; details are omitted in the interest of brevity. ■

Lemma A.2 *Consider slowly varying functions $u, w : \mathbb{R}_+ \to \mathbb{R}_+$, such that $u(x) \sim w(x)$ $(x \to \infty)$, and let $\alpha > 1$. For any sequences $\{\zeta_r,\ r = 1, 2, \ldots\}$ and $\{\eta_r,\ r = 1, 2, \ldots\}$ with $\lim_{r\to\infty} \zeta_r = \lim_{r\to\infty} \eta_r = \infty$ such that*

$$\lim_{r\to\infty} r\zeta_r^{-\alpha} u(\zeta_r) = \lim_{r\to\infty} r\eta_r^{-\alpha} w(\eta_r) = K \tag{A.5}$$

for some finite constant $K > 0$, it holds that $\zeta_r \sim \eta_r$ $(r \to \infty)$.

Proof. We first look at the special case when $u = w$, in which case condition (A.5) implies

$$\lim_{r\to\infty} \frac{\zeta_r^{-\alpha} u(\zeta_r)}{\eta_r^{-\alpha} u(\eta_r)} = 1. \tag{A.6}$$

We refer to the proof of Lemma 3.8.1, where we introduced the asymptotically equivalent representation (3.66) of the slowly varying function u. Substituting (3.66) in (A.6), we see that

$$\lim_{r\to\infty} \exp\left(-\int_A^{\zeta_r} \frac{\alpha - \varepsilon(t)}{t}\, dt + \int_A^{\eta_r} \frac{\alpha - \varepsilon(t)}{t}\, dt \right) = 1,$$

or, equivalently,

$$\lim_{r\to\infty} \left| \int_{\zeta_r}^{\eta_r} \frac{\alpha - \varepsilon(t)}{t}\, dt \right| = 0. \tag{A.7}$$

Pick δ in $(0, \alpha)$. Because $\lim_{r\to\infty} \zeta_r = \lim_{r\to\infty} \eta_r = \infty$, there exists r_δ such that for $r > r_\delta$ we have $|\varepsilon(t)| < \delta$ whenever $t > \min(\zeta_r, \eta_r)$. Thus,

$$(\alpha - \delta) \left| \ln \frac{\eta_r}{\zeta_r} \right| < \left| \int_{\zeta_r}^{\eta_r} \frac{\alpha - \varepsilon(t)}{t}\, dt \right|, \quad r > r_\delta,$$

and combining this last inequality with (A.7) we obtain the desired conclusion $\lim_{r\to\infty} \eta_r/\zeta_r = 1$.

In general, when u and w are not necessarily equal, we note the easy relation

$$\frac{\zeta_r^{-\alpha} u(\zeta_r)}{\eta_r^{-\alpha} u(\eta_r)} = \frac{r\zeta_r^{-\alpha} u(\zeta_r)}{r\eta_r^{-\alpha} w(\eta_r)} \cdot \frac{w(\eta_r)}{u(\eta_r)}, \quad r = 1, 2, \ldots$$

Condition (A.5) and the asymptotic equivalence of u and w together imply that the relation (A.6) still holds, and the conclusion $\zeta_r \sim \eta_r$ $(r \to \infty)$ follows from the first part of the proof. ■

Proposition A.3 *Consider a slowly varying function $u : \mathbb{R}_+ \to \mathbb{R}_+$, and let $\alpha > 1$. For any sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ with $\lim_{r\to\infty} \zeta_r = \infty$ such that (A.5) holds, we have*

$$\zeta_r \sim r^{\frac{1}{\alpha}} w(r) \quad (r \to \infty)$$

for some slowly varying function $w : \mathbb{R}_+ \to \mathbb{R}_+$.

Proof. We go back to the proof of Lemma 3.8.1, where we introduced the asymptotically equivalent representation (3.66) of the slowly varying function u. In view of Lemma A.2, it suffices to consider a sequence $\{\zeta_r,\ r = 1, 2, \ldots\}$ determined by the relations

$$r\zeta_r^{-\alpha} \cdot c \exp\left(\int_A^{\zeta_r} \frac{\varepsilon(t)}{t}\, dt\right) = K, \quad r \geq r_\star \tag{A.8}$$

for some $r^\star$ large enough, with constants $A > 0$ and $c > 0$, and Borel mapping $\varepsilon : \mathbb{R}_+ \to \mathbb{R}$ such that $\lim_{t\to\infty} \varepsilon(t) = 0$. We can write (A.8) in the equivalent form

$$Br^{\frac{1}{\alpha}} = \exp\left(\int_A^{\zeta_r} \frac{1 - \beta(t)}{t}\, dt\right), \quad r \geq r_\star$$

with

$$B := \left(\frac{c}{KA^\alpha}\right)^{\frac{1}{\alpha}} \quad \text{and} \quad \beta(t) := \frac{1}{\alpha}\varepsilon(t), \quad t \geq 0.$$

Hence, by Proposition A.1, for large enough r we see that ζ_r is the unique solution $x(y)$ of the equation (A.1) with $y = Br^{\frac{1}{\alpha}}$. By the second part of Proposition A.1,

we have

$$\zeta_r = x(Br^{\frac{1}{\alpha}}) \sim Br^{\frac{1}{\alpha}} v(Br^{\frac{1}{\alpha}}) \quad (r \to \infty)$$

for some slowly varying function $v : \mathbb{R}_+ \to \mathbb{R}_+$. The desired conclusion is now immediate once we note that the mapping $w : x \to Bv(Bx^{\frac{1}{\alpha}})$ is slowly varying whenever v is. ■

Appendix B

Stochastic orderings

We collect here some definitions and properties concerning stochastic orderings. The material is drawn mostly from [59], where additional information is available.

Throughout let X and Y denote two $\mathbb{R}$–valued rvs, and let $\mathcal{D}$ denote the set of all probability distribution functions of $\mathbb{R}$–valued rvs.

Definition B.1 *Let X and Y have distribution functions F and G, respectively. We say that X is stochastically smaller than Y, and write $X \leq_{st} Y$, or, equivalently, $F \leq_{st} G$, if*

$$F(x) \geq G(x), \quad x \in \mathbb{R}\,.$$

Proposition B.2 *It holds that $X \leq_{st} Y$ if and only if*

$$\mathbf{E}\left[\phi(X)\right] \leq \mathbf{E}\left[\phi(Y)\right] \tag{B.1}$$

for all non-decreasing functions $\phi : \mathbb{R} \to \mathbb{R}$ for which the expectations in (B.1) exist.

Proposition B.3 *Let $\{F_n,\ n = 1, 2, \ldots\}$ and $\{G_n,\ n = 1, 2, \ldots\}$ be two subsets of $\mathcal{D}$ such that $F_n \Longrightarrow_n F$ and $G_n \Longrightarrow_n G$ for limits F and G in $\mathcal{D}$, respectively. If $F_n \leq_{st} G_n$ for all $n = 1, 2, \ldots$ then $F \leq_{st} G$.*

Definition B.4 *We say that X is smaller in the convex stochastic ordering than Y, and write $X \leq_{cx} Y$, if*

$$\mathbf{E}\left[\phi(X)\right] \leq \mathbf{E}\left[\phi(Y)\right] \tag{B.2}$$

for all convex functions $\phi : \mathbb{R} \to \mathbb{R}$ whenever the expectations exist in (B.2).

We similarly define the increasing convex stochastic ordering.

Definition B.5 *We say that X is smaller in the increasing convex stochastic ordering than Y, and write $X \leq_{icx} Y$, if*

$$\mathbf{E}\left[\phi(X)\right] \leq \mathbf{E}\left[\phi(Y)\right] \tag{B.3}$$

for all increasing convex functions $\phi : \mathbb{R} \to \mathbb{R}$ whenever the expectations exist in (B.3).

The $\leq_{icx}$ ordering admits the following characterization [59, p. 8].

Proposition B.6 *It holds that $X \leq_{icx} Y$ if and only if X is smaller in mean residual life than Y, i.e.,*

$$\mathbf{E}\left[(X-x)^+\right] \leq \mathbf{E}\left[(Y-x)^+\right], \quad x \in \mathbb{R}$$

provided the expectations above are finite.

Consider now the Lindley recursions

$$w_{n+1}^{(X)} = [w_n^{(X)} + X_{n+1}]^+ \quad \text{and} \quad w_{n+1}^{(Y)} = [w_n^{(Y)} + Y_{n+1}]^+ \quad n = 0, 1, \ldots, \tag{B.4}$$

with initial conditions $w_0^{(X)}$ and $w_0^{(Y)}$; these are independent of the driving sequences of i.i.d. rvs $\{X_n,\ n = 1, 2, \ldots\}$ and $\{Y_n,\ n = 1, 2, \ldots\}$, with generic rvs X and Y, respectively. Let $\prec$ denote either $\leq_{st}$ or $\leq_{icx}$, and let $\{w_n,\ n = 0, 1, \ldots\}$ be either $\{w_n^{(X)},\ n = 0, 1, \ldots\}$ or $\{w_n^{(Y)},\ n = 0, 1, \ldots\}$.

Proposition B.7 [59, p. 79] *If $w_0 \prec w_1$ in (B.4) then for all $n = 0, 1, \ldots$, we have*

$$w_n \prec w_{n+1}.$$

Moreover, if the stationary rv w_∞ exists, then

$$w_n \prec w_\infty, \quad n = 0, 1, \ldots,$$

where, if $\prec$ denotes $\leq_{icx}$, it is further assumed that w_∞ has finite expectation.

Proposition B.8 [59, p. 80] *If $X \prec Y$ and $w_0^{(X)} \prec w_0^{(Y)}$ in (B.4) then*

$$w_{n+1}^{(X)} \prec w_{n+1}^{(Y)}, \quad n = 0, 1, \ldots.$$

Moreover, if the corresponding stationary versions $w_\infty^{(X)}$ and $w_\infty^{(Y)}$ exist, then $X \prec Y$ is sufficient to ensure that

$$w_\infty^{(X)} \prec w_\infty^{(Y)}$$

where, if $\prec$ denotes $\leq_{icx}$, it is further assumed that $w_\infty^{(Y)}$ has finite expectation.

Bibliography

[1] R. G. Addie, M. Zukerman, and T. Neame. Fractal traffic: Measurements, modeling and performance evaluation. In *IEEE Infocom 95*, pages 985–992, Boston (MA), April 1995.

[2] F. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm–Martingale Calculus and Stochastic Recurrences*, volume 26 of *Applications of Mathematics*. Springer–Verlag, Berlin Heidelberg, 1994.

[3] J. Beran. *Statistics for Long-Memory Processes*. Chapman and Hall, New York (NY), 1994.

[4] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long–range dependence in variable bit–rate video traffic. *IEEE Transactions on Communications*, 43:1566–1579, 1995.

[5] P. Billingsley. *Convergence of Probability Measures*. John Wiley and Sons, 1968.

[6] N. H. Bingham. Fluctuation theory in continuous time. *Advances in Applied Probability*, 7:705–766, 1975.

[7] N. H. Bingham, C. M. Goldie, and J. T. Teugels. *Regular Variation.* Encyclopedia of Mathematics and its Applications. Cambridge Univerity Press, Cambridge (UK), 1987.

[8] O. J. Boxma and J. W. Cohen. Heavy traffic analysis for the GI/GI/1 queue with heavy tailed distributions. Technical Report PNA-R9710, CWI, Amsterdam, 1997.

[9] F. Brichet, J. Roberts, A. Simonian, and D. Veitch. Heavy traffic analysis of a storage model with long range dependent on/off sources. *Queueing Systems – Theory and Applications*, 23:197–215, 1996.

[10] J. W. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic Processes and their Applications*, 2:31–58, 1974.

[11] D. R. Cox. Long–range dependence: A review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, pages 55–74. The Iowa State University Press, Ames (IA), 1984.

[12] M. E. Crovella and A. Bestavros. Self–similarity in World Wide Web traffic: Evidence and possible causes. In *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, May 1996.

[13] G. Doetsch. *Guide to the Applications of Laplace Transforms.* D. Van Nostrand Company Ltd., London (UK), 1963.

[14] N. G. Duffield. On the relevance of long–tailed durations for the statistical multiplexing of large aggregations. In *34th Annual Allerton Conference on Communications, Control and Computing*, October 1996.

[15] N. G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 118, pages 363–374, 1995.

[16] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events*. Applications of Mathematics. Springer–Verlag, 1997.

[17] A. Erdélyi. *Higher Transcendental Functions*, volume 3. McGraw-Hill, New York (NY), 1955.

[18] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long–range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4:209–223, 1996.

[19] J. D. Esary, F. Proschan, and D. W. Walkup. Association of random variables, with applications. *Annals of Mathematical Statistics*, 38:1466–1474, 1967.

[20] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley and Sons, New York (NY), 1972.

[21] P. J. Fleming and B. Simon. Interpolation approximations of sojourn time distributions. *Operations Research*, 39(2):251–260, 1991.

[22] H. J. Fowler and W. E. Leland. Local area network traffic characteristics, with implications for broadband network congestion management. *IEEE Journal on Selected Areas in Communications*, 9:1139–1149, 1991.

[23] M. Garrett and W. Willinger. Analysis, modeling and generation of self–similar VBR video traffic. In *ACM SIGCOMM 94*, pages 269–280, September 1994.

[24] M. Grossglauser and J.-C. Bolot. On the relevance of long–range dependence in network traffic. *IEEE/ACM Transactions on Networking*. To appear.

[25] P. Hall and C. C. Heyde. *Martingale Limit Theory and its Applications*. Academic Press, New York (NY), 1980.

[26] J. M. Harrison. The supremum distribution of a Lévy process with no negative jumps. *Advances in Applied Probability*, 9:417–422, 1977.

[27] J. M. Harrison. *Brownian Motion and Stochastic Flow Systems*. John Wiley and Sons, New York (NY), 1985.

[28] D. Heath, S. Resnick, and G. Samorodnitsky. Heavy tails and long range dependence in on/off processes and associated fluid models. Technical report, Cornell University, 1996.

[29] P. R. Jelenković and A. A. Lazar. Multiplexing on–off sources with subexponential on periods: Part I. In *IEEE Infocom 97*, Kobe (Japan), April 1997.

[30] P. R. Jelenković, A. A. Lazar, and N. Semret. The effect of multiple time scales and subexponentiality in MPEG video streams on queueing behavior. *IEEE Journal on Selected Areas in Communications*, 15(6):1052–1071, August 1997.

[31] J. F. C. Kingman. The single server queue in heavy traffic. *Proceedings of the Cambridge Philosophical Society*, 57:902–904, 1961.

[32] J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society*, 24:383–392, 1962.

[33] T. Konstantopoulos and S.-J. Lin. Fractional Brownian motions and Lévy motions as limits of stochastic traffic models. In *34th Annual Allerton Conference on Communication, Control and Computation*, pages 913–922, October 1996.

[34] M. M. Krunz and A. M. Makowski. Modeling video traffic using $M|G|\infty$ input processes: A compromise between Markovian and LRD models. *IEEE Journal on Selected Areas in Communications*, 16(5):733–748, June 1998.

[35] J. Lamperti. Semi–stable stochastic processes. *Transactions of the American Mathematical Society*, 104:62–78, 1962.

[36] W. Leland, M. S. Taqqu, W. Willinger, and D. Wilson. On the self–similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, 1994.

[37] N. Likhanov, B. Tsybakov, and N. D. Georganas. Analysis of an ATM buffer with self–similar (fractal) input traffic. In *IEEE Infocom 95*, pages 985–992, Boston (MA), April 1995.

[38] G. C. Lin and T. Suda. On the impact of long–range dependent traffic in dimensioning ATM network buffer. In *IEEE Infocom 98*, pages 1317–1324, April 1998.

[39] Z. Liu, Ph. Nain, D. Towsley, and Z.-L. Zhang. Asymptotic behavior of a multiplexer fed by a long–range dependent process. *Journal of Applied Probability*. To appear.

[40] M. Livny, B. Melamed, and A. K. Tsiolis. The impact of autocorrelation on queueing systems. *Management Science*, 39(3):322–339, March 1993.

[41] B. B. Mandelbrot and J. W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

[42] C. M. Newman and A. L. Wright. An invariance principle for certain dependent sequences. *The Annals of Probability*, 9:671–675, 1981.

[43] I. Norros. A storage model with self–similar input. *Queueing Systems – Theory and Applications*, 16:387–396, 1994.

[44] I. Norros. On the use of fractional Brownian motion in the theory of connectionless networks. *Journal on Selected Areas in Communications*, 13(6):953–962, August 1995.

[45] M. Parulekar and A. M. Makowski. Tail probabilities for a multiplexer with self–similar traffic. In *IEEE Infocom 96*, San Francisco (CA), April 1996.

[46] M. Parulekar and A. M. Makowski. $M|G|\infty$ input processes : A versatile class of models for network traffic. In *IEEE Infocom 97*, Kobe (Japan), April 1997.

[47] M. Parulekar and A. M. Makowski. Tail probabilities for $M|G|\infty$ processes (I): Preliminary asymptotics. *Queueing Systems – Theory and Applications*, 27:271–296, 1997.

[48] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1993.

[49] M. I. Reiman and B. Simon. An interpolation approximation for queueing systems with Poisson input. *Operations Research*, 36:454–469, 1988.

[50] M. I. Reiman and B. Simon. Light traffic limits of sojourn time distributions in Markovian queueing networks. *Stochastic Models*, 4:191–233, 1988.

[51] M. I. Reiman and B. Simon. Open queueing systems in light traffic. *Mathematics of Operations Research*, 14:26–59, 1989.

[52] B. K. Ryu and A. Elwalid. The importance of long–range dependence of VBR video traffic in ATM traffic engineering: Myths and realities. In *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, May 1996.

[53] G. Samorodnitsky and M. S. Taqqu. *Stable non–Gaussian Random Processes*. Chapman and Hall, London, 1994.

[54] K. Sigman. *Stationary marked point processes: An intuitive approach*. Chapman and Hall, New York (NY), 1995.

[55] K. Sigman and G. Yamazaki. Fluid models with burst arrivals: A sample path analysis. *Probability in the Engineering and Informational Sciences*, 6:17–27, 1992.

[56] B. Simon. A simple relationship between light and heavy traffic limits. *Operations Research*, 40:S342–S345, 1992.

[57] A. V. Skorokhod. Limit theorems for stochastic processes with independent increments. *Theory of Probability and its Applications*, 2:138–171, 1957.

[58] V. Solo. On queueing theory for broadband communication network traffic with long range correlation. In *Proceedings of the 34th Conference on Decision and Control*, pages 853–858, New Orleans (LA), December 1995.

[59] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley and Sons, New York (NY), 1984.

[60] K. P. Tsoukatos and A. M. Makowski. Heavy traffic analysis for a multiplexer driven by $M|G|\infty$ input processes. Technical Report 96–70, Institute for Systems Research, University of Maryland, College Park, September 1996.

[61] K. P. Tsoukatos and A. M. Makowski. Heavy traffic analysis for a multiplexer driven by $M|G|\infty$ input processes. In V. Ramaswami and P. E. Wirth, editors, *15th International Teletraffic Congress*, pages 497–506, Washington (DC), June 1997. Top–5 finisher in student paper contest.

[62] K. P. Tsoukatos and A. M. Makowski. Heavy traffic limits associated with $M|G|\infty$ input processes. *Queueing Systems – Theory and Applications*, 1999. To appear.

[63] K. P. Tsoukatos and A. M. Makowski. Interpolation approximations for $M|G|\infty$ arrival processes. In *IEEE ICC 99*, Vancouver (BC, Canada), June 1999. To appear.

[64] B. von Bahr and C. G. Esseen. Inequalities for the r^{th} absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Annals of Mathematical Statistics*, 36:299–303, 1965.

[65] W. Whitt. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, 5:67–85, 1980.